Essays on Instrumental Variables

Essays on Instrumental Variables

A dissertation presented

by

Michal Kolesar

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

April 2013

Dissertation Advisors:
Professor Guido W. Imbens
Professor Gary Chamberlain

Author:
Michal Kolesar

Essays on Instrumental Variables

Abstract

This dissertation addresses issues that arise in the classic linear instrumental variables (IV) model when some of the underlying assumptions are violated.

Chapter 1 analyzes estimators based on this model when the treatment effects are heterogeneous. I show if the local average treatment effects vary, two-step IV estimators like the two-stage least squares (TSLS) estimator typically all estimate the same convex combination of them. In contrast, the estimand of the limited information maximum likelihood (LIML) estimator may be outside of the convex hull of the local average treatment effects. This result questions the standard recommendation to use LIML when the number of instruments is large as a way of addressing the bias exhibited by TSLS in these settings. Instead, I propose a new estimator, a version of the jackknife instrumental variables estimator (UJIVE). Unlike TSLS or LIML, UJIVE is consistent for a convex combination of local average treatment effects under many instrument asymptotics that also allow for many covariates and heteroscedasticity.

Chapter 2 studies estimation and inference when the instruments are have direct effects on the outcome, and thus are "invalid". The novel identifying assumption is that the direct effects of these invalid instruments are uncorrelated with their effects on the endogenous regressor. Under this assumption, LIML is no longer consistent, but a modification of the bias-corrected TSLS estimator remains consistent.

Chapter 3 derives a principled and unified approach to inference when number of instruments is large. I use an invariance property of the model to construct an integrated

likelihood which by design yields inference procedures that are valid under many instrument asymptotics and asymptotically optimal under rotation invariance and Gaussian errors. I establish that this integrated likelihood coincides with the random-effects likelihood of Chamberlain and Imbens (2004), and that the maximum likelihood estimator of the parameter of interest coincides with the LIML estimator. The random effects framework can be generalized to allow the instruments to have direct effects on the outcome, as in Chapter 2. The resulting maximum likelihood estimator is a mixture between the bias-corrected TSLS estimator and LIML.

Contents

Acknowledgments

I will always be deeply grateful to Guido Imbens and Gary Chamberlain, my thesis advisors, for their constant guidance, support, and encouragement. Their ideas and suggestions have been invaluable in making each of the chapters of this thesis possible. Working with them has been a truly stimulating intellectual experience. Raj Chetty taught me a lot about applying econometric tools in practice. I am grateful to Jim Stock for many helpful conversations and suggestions. Chapter 2 of this thesis has been co-written with Raj Chetty, John Friedman, Edward Glaeser, and Guido Imbens; I thank them for permission to include this joint research in my thesis.

I also received helpful comments from Alberto Abadie, Nikhil Agarwal, Josh Angrist, Paul Goldsmith-Pinkham, Adam Guren, Pepe Montiel, Whitney Newey, Fanyin Zheng, and participants in the econometrics lunch and labor seminars at Harvard University, the Harvard-MIT Econometrics seminar, and the Oslo 2011 Econometric Society meeting.

The Department of Economics at Harvard University provided generous financial support while I was working on these chapters.

Introduction

This dissertation focuses on identification and estimation in instrumental variables (IV) models. Estimators typically used in practice, such as the two-stage least squares (TSLS) estimator or the limited information maximum likelihood (LIML) estimator, are based on the classic linear instrumental variables model. This model imposes substantive restrictions one the relationship between the outcome, the endogenous variable, and the instruments, which may not be satisfied in many empirical settings. The common goal of the three chapters in my dissertation is to analyze the properties of the classic linear IV estimators when these restrictions are weakened, as well as develop new, more robust estimators that work well under these weaker restrictions.

Estimation in an instrumental variables model with treatment effect heterogeneity

The classic linear instrumental variables model is often used to estimate the causal effect of the endogenous variable (treatment) on the outcome. When the individual treatment effect is independent of treatment status and covariates, estimators based on this model estimate the population average treatment effect (Heckman, 1997). However, since this assumption rules out selection into treatment based on anticipated gains from treatment, it is not very plausible in many empirical settings.

In Chapter 1, I analyze these estimators when the heterogeneity in treatment effects is unrestricted, as in Imbens and Angrist (1994). I divide the estimators into two classes:

two-step instrumental variables (TSIV) estimators that include the two-stage least squares (TSLS) estimator; and minimum distance estimators that include the limited information maximum likelihood (LIML) estimator. I show that if the local average treatment effects vary, estimators in the TSIV class typically all estimate the same convex combination of them. In contrast, estimands of minimum distance estimators may be outside of the convex hull of the local average treatment effects, and may therefore not correspond to a causal effect.

This result questions the standard recommendation to use LIML when the number of instruments is large as a way of addressing the bias exhibited by TSLS in these settings. Instead, I propose a new TSIV estimator, the unbiased jackknife instrumental variables estimator (UJIVE). This estimator is similar to the jackknife instrumental variables estimator (JIVE, Phillips and Hale, 1977; Angrist, Imbens and Krueger, 1999) in that it also uses a "leave-one-out" jackknife-type predictor of the treatment in the first stage. Unlike JIVE, however, UJIVE also uses the "leave-one-out" predictor to partial out the effect of the covariates. This ensures that the single constructed instrument is uncorrelated with the outcome even in finite samples. Consequently, unlike TSLS JIVE, or LIML, UJIVE is consistent for a convex combination of local average treatment effects under many instrument asymptotics that also allow for many covariates and heteroscedasticity. I therefore recommend that in settings with many instruments researchers use UJIVE, instead of TSLS or LIML.

Identification and Inference with Many Invalid Instruments

In chapter 2, written jointly with Raj Chetty, John Friedman, Edward Glaeser, and Guido Imbens, we study estimation and inference in settings where the interest is in the effect of a potentially endogenous regressor on some outcome. To address the endogeneity, we exploit the presence of additional variables. Like conventional instrumental variables, these variables are correlated with the endogenous regressor. However, unlike conventional instrumental variables, they also have direct effects on the outcome, and thus are "invalid" instruments. Our novel identifying assumption is that the direct effects of these invalid instruments are uncorrelated with the effects of the instruments on the endogenous regressor.

To motivate this assumption, suppose that we are interested in estimating the effect of early achievement for children, as measured by kindergarten performance, on subsequent outcomes, say first grade scores, as in Chetty, Friedman, Hilger, Saez, Schanzenbach and Yagan (2011). We want to exploit the fact that in the in Tennessee's Student/Teacher Achievement Ratio (STAR) Project, teachers were randomly assigned to kindergarten classes, and so we use classroom indicators as instruments for kindergarten performance. Suppose that kindergarten teachers only affect first-grade scores through their effect on kindergarten scores, so that the instrument is valid in this sense. However, since classes mostly stay together in subsequent years, kindergarten teacher assignment will be perfectly correlated with first grade teacher assignment. Therefore, the instrument (kindergarten classroom assignment) may have direct effects on the outcome (first grade performance) through first grade classroom assignment, that is not mediated through the endogenous regressor (kindergarten performance). Yet if first grade teachers are also randomly assigned, and thus independent of kindergarten teacher assignment, the direct effect of the instrument on the outcome might reasonably be assumed to be uncorrelated with the direct effect of the instrument on the endogenous regressor.

We show that under our new identifying assumption, LIML is no longer consistent, but that a modification of the bias-corrected TSLS estimator remains consistent. We also show that conventional tests for over-identifying restrictions, adapted to the many instruments setting, can be used to test for the presence of these direct effects. We recommend that empirical researchers carry out such tests and compare estimates based on LIML and the modified version of bias-corrected TSLS. We illustrate in the context of two applications that such practice can be illuminating, and that our novel identifying assumption has substantive empirical content.

Random-effects approach to inference with many instruments

In Chapter 3, I provide a principled and unified way of doing inference in a linear instrumental variables model with homoscedastic errors in which the number of instruments

is potentially large. The presence of a large number of instruments creates an incidental parameter problem (Neyman and Scott, 1948) because the number of first-stage coefficients corresponds to the number of instruments. To directly address the incidental parameter problem, I use an invariance property of the model and a Bernstein-von Mises type argument to construct an integrated likelihood, which by design yields inference procedures that are valid under many instrument asymptotics and asymptotically optimal under rotation invariance. I establish that this integrated likelihood coincides with the random-effects likelihood of Chamberlain and Imbens (2004), and that the maximum likelihood estimator of the parameter of interest coincides with LIML.

This analysis yields new insights into the sources of identification in the instrumental variables model, and I use these insights to relax the basic setup along two dimensions. First, I use it to derive an estimator that is more efficient than LIML when the assumption that the errors are Normally distributed is dropped. In particular, maximizing the random-effects likelihood is equivalent to minimizing a minimum distance objective function with respect to a particular weight matrix. This weight matrix is optimal if the errors in the instrumental variables model are Normally distributed, but not otherwise; using weights proportional the inverse of the asymptotic covariance matrix of the moment conditions yields a more efficient estimator.

Second, I relax the exclusion restriction by allowing the instruments to have direct effects on the outcome. As long as these direct effects are orthogonal to the effects of the instruments on the endogenous regressor, the coefficient on the endogenous regressor is still identified. I generalize the random effects likelihood to allow for such effects, and show that the resulting maximum likelihood estimator is a mixture between the bias-corrected two-stage least squares estimator and LIML.

Chapter 1

Estimation in an instrumental variables model with treatment effect heterogeneity

1.1 Introduction

The classic linear instrumental variables model is commonly used to estimate treatment effects. When the individual treatment effect is independent of treatment status and covariates, estimators based on this model estimate the population average treatment effect (Heckman, 1997). However, since this assumption rules out selection into treatment based on anticipated gains from treatment, it is not very plausible in many empirical settings. It is therefore important to understand the properties of these estimators when the individual treatment effect is allowed to be correlated with treatment status.

The first contribution of this chapter is to characterize the estimands of estimators based on the classic linear instrumental variables (IV) model when the treatment effects are unrestricted. I assume that the instruments satisfy the monotonicity condition of Imbens and Angrist (1994), so that for each pair of instrument values, we can identify a local average treatment effect (LATE). I show that the two-stage least squares (TSLS) estimator, under

some mild assumptions about the first stage, estimates a convex combination of these local average treatment effects, weighted over different pairs of instrument values and covariates. On the other hand, unless all LATEs are the same, the estimand of the limited information maximum likelihood (LIML, Anderson and Rubin, 1949) depends on the covariance matrix of the reduced-form errors, and may lie outside the convex hull of the local average treatment effects. Therefore, the estimand may not correspond to a causal effect. Moreover, other estimators based on the classic linear IV model will, depending on how they are constructed, either estimate the same convex combination of LATEs as TSLS, or else behave similarly to LIML.

In particular, estimators that behave like TSLS can be thought of as two-step estimators. In the first step, they construct a single instrument, a predictor of the treatment status based on the first-stage regression. In the second step, an instrumental variables estimator that uses this constructed instrument as a single instrument is used to estimate the treatment effect. I refer to these estimators as two-step instrumental variables estimators. In the limit under standard asymptotics, the exact way of constructing the single instrument does not matter; all two-step IV estimators converge to the same probability limit as the infeasible instrumental variables estimator that uses a population linear predictor of the treatment status as a single instrument. In turn, the probability limit of this IV estimator corresponds to a weighted average of LATEs. The weights are non-negative if the single instrument itself satisfies monotonicity in that changing its value does not induce two-way flows in and out of treatment.

In contrast, estimators that behave like LIML are based on the property of the classic linear IV model that the coefficients on the instruments in the first-stage regression are proportional to the coefficients in the reduced-form outcome regression. These estimators, which I refer to as minimum distance estimators, minimize a minimum distance objective function that directly enforces this proportionality with respect to some weight matrix. The estimator of the treatment effect is given by the estimator of the constant of proportionality. Goldberger and Olkin (1971) show that LIML can be thought of in this way, with the weight

matrix depending on the covariance matrix of the reduced-form errors.

This approach yields a different estimand under treatment effect heterogeneity because imposing proportionality of the reduced-form coefficients implies that the treatment and the outcome are treated symmetrically. In particular, it requires that the estimand of the reverse two-stage least squares estimator (RTSLS) be equal to the estimand of TSLS. The RTSLS estimator is obtained as the reciprocal of the TSLS estimator in the instrumental variables model that swaps the treatment and the outcome. This requirement makes sense if the instrumental variables model is supposed to solve an errors-in-variables problem (Zellner, 1970), or an omitted variable bias (Chamberlain, 2007). However, in the context of estimating treatment effects, the reduced-form coefficients are no longer proportional to each other unless all LATEs are equal. Therefore, the TSLS and RTSLS estimands are in general different; the probability limit of the RTSLS estimator is the same as that of an instrumental variables estimator that uses a linear predictor of the outcome based on the reduced-form outcome regression as an instrument. This instrument induces a different weighting scheme for the LATEs, and hence a different estimand, than using a linear predictor of the treatment status as an instrument. Unlike the TSLS weights, these weights are proportional to the effect size, with the bigger LATEs receiving more weight.

There are two ways in which this difference between TSLS and RTSLS estimands can cause a minimum distance estimand to be outside the convex hull of LATEs. First, if some LATEs are negative, the RTSLS estimand gives them a negative weight, so that the estimand may end up being outside the convex hull of the LATEs. Consequently, the minimum distance estimand, trying to equate RTSLS with TSLS, may end up being outside the convex hull. Second, even if the RTSLS estimand is inside the convex hull, if the weight matrix that is used to equate RTSLS with TSLS is non-diagonal, as is the case with LIML, the minimum distance estimand is not guaranteed to lie between the RTSLS and TSLS estimands.

In settings with a few strong instruments, it is easy to avoid these problems by simply avoiding LIML and using TSLS. However, when many instruments are used, TSLS may be severely biased even in large samples (Bound, Jaeger and Baker, 1995), and it is inconsistent

under the many instrument asymptotic sequence of Kunitomo (1980), Morimune (1983), and Bekker (1994). Therefore, when the number of instruments is large, the standard recommendation has been to use LIML, which is not only consistent under many instrument asymptotics, but also efficient among rotation invariant estimators and homoscedasticity (Chioda and Jansson, 2009; Anderson, Kunitomo and Matsushita, 2010). Recently, other estimators have been proposed that behave better than LIML under heteroscedasticity. Hausman, Newey, Woutersen, Chao and Swanson (2012) propose a Fuller (1977) type modification to a jackknife version of LIML (HLIM). Bekker and Crudu (2012) propose a similar estimator, which they call symmetric jackknife. However, all of these estimators are minimum distance estimators, and therefore not likely to work well under treatment effect heterogeneity.

The second contribution of this chapter is to propose a new estimator in the two-step IV class, the unbiased jackknife IV estimator (UJIVE), that remains consistent for a convex combination of LATEs even under many instrument asymptotics and heteroscedasticity. This estimator is similar to the jackknife instrumental variables estimator (JIVE, Phillips and Hale, 1977; Angrist *et al.*, 1999) in that it also uses a "leave-one-out" jackknife-type predictor of the treatment in the first stage, but differs from JIVE in the way it deals with covariates. In particular, in constructing the single instrument in the two-step IV procedure, we need to partial out the effect of covariates on the treatment. Suppose, for example, that the instruments are classroom indicators, and the covariates are school indicators (school "fixed effects"). Then the JIVE estimate of the effect of covariates on the treatment status of individual i is given by an average treatment status of individuals in the same school as individual i. With a finite number of observations in each school, this estimate is noisy, and since it depends on the treatment status of individual i, the estimation error is correlated with the outcome. Therefore, the single constructed instrument is also correlated with the outcome, causing JIVE to be biased when the number of covariates is large (Ackerberg and Devereux, 2009). In contrast, the UJIVE estimate of the effect of the covariates is given by a sample average that excludes individual i, which guarantees that the prediction error will

be uncorrelated with the outcome. As a result, unlike JIVE, UJIVE is consistent for a convex combination of LATEs even when we let the number of covariates, in addition to the number of instruments, increase in proportion to the sample size, as in Anatolyev (2011) and Kolesár, Chetty, Friedman, Glaeser and Imbens (2011).

The estimand of two-step IV estimators can be seen as one way of summarizing the effect of the treatment on outcome. For particular policy questions, however, we might be interested in a weighting scheme that is different than the one used by these estimators. For this purpose, a number of alternative approaches, not based on the classic IV model, have been proposed in the literature. For example, Frölich (2007) derives a non-parametric estimator for the largest subpopulation of compliers for which a treatment effect can be identified. When the instrument is binary, Abadie (2003) works out a semi-parametric approach to approximating a treatment response function, and Hirano, Imbens, Rubin and Zhou (2000) and Yau and Little (2001) use a parametric approach to estimate a LATE that does not condition on covariates. To keep the chapter focused, I do not try to compare the classic IV estimators with these alternative approaches.

The rest of the chapter is organized as follows. In Section 1.2, I set up the problem of estimating causal effects in a potential outcomes framework. In Section 1.3, I review assumptions underlying the classic linear IV model, and I introduce the classes of two-step IV and minimum distance estimators that are based on this model. In Section 1.4, I introduce the local average treatment effects framework of Imbens and Angrist (1994). In Section 1.5, I derive the first main result of the chapter, the estimands of two-step IV and minimum distance estimators under the LATE assumptions. In Section 1.6, I derive the second main result of the chapter that UJIVE is consistent for a convex combination of LATEs under many instrument asymptotics. Section 1.7 concludes. Proofs are collected in Appendix A.

1.2 Potential outcomes framework

We want to learn about the causal effect of a treatment on some outcome of interest using a random sample of n individuals indexed by $i = 1, \ldots, n$. For clarity of exposition, I focus on

the case when the treatment is binary. Let T_i be an indicator for receiving treatment, so that $T_i = 1$ if individual i receives treatment and zero otherwise. Let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes in the treated and untreated states. The treatment effect for individual i is then given by $\tau_i = Y_i(1) - Y_i(0)$.

The fundamental problem is that for each individual, we only observe the potential outcome corresponding to the observed treatment state, $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$; the other potential outcome is not observed. Therefore, we cannot compute τ_i directly for any individual. Moreover, there is a concern that anticipated potential outcomes affect selection into treatment, so that comparing the average outcome of the subsample of individuals who are treated in our sample with those who are not is likely to lead to a biased estimate of the population average treatment effect $\mathbb{E}[\tau_i]$.

We do, however, observe instruments Q_i with support $\mathcal{Q}$ that help to identify average treatment effects for at least some subpopulations. Following the notation in Imbens and Angrist (1994), for each possible realization $q \in \mathcal{Q}$, let $T_i(q)$ denote the potential treatment variable that equals one if individual i would receive treatment if their instrument value was changed to $Q_i = q$, and equals zero if they would not receive treatment. The observed treatment status is given by $T_i = T_i(Q_i)$; the other potential treatments are not observed.

We also observe a vector of covariates X_i with support $\mathcal{X}$. I include these covariates explicitly for two related reasons. First, in many empirical applications the identification assumptions that underlie the instrumental variables framework may only be plausible conditional on X_i. One simple approach in this case is to carry out the analysis separately for all values of the covariates. However, when the covariate set is detailed so that the support $\mathcal{X}$ is rich, this approach is unlikely to be satisfactory. Second, even when the identification assumptions are plausible unconditionally, inference without covariates might not be precise enough. A common solution to both of these problems in practice is to estimate a single model with covariates. It is therefore important to understand how the presence of covariates affects inference.

In summary, the observed data vector for each individual is given by (Y_i, T_i, Q_i, X_i).

For simplicity I will assume that the support of the observed data vector is given by the Cartesian product $\mathbb{R} \times \{0,1\} \times \mathcal{Q} \times \mathcal{X}$. This will ensure that we can freely manipulate T_i and Q_i while keeping X_i constant, so that the set of potential outcomes and treatments $\{Y_i(t), T_i(q)\}_{t\in\{0,1\},q\in\mathcal{Q}}$ is well-defined.

Two important functions of the distribution of the observed data are given by the two regression functions

$$r(q,x) = \mathbb{E}[Y_i \mid Q_i = q, X_i = x], \tag{1.1}$$

$$p(q,x) = \mathbb{E}[T_i \mid Q_i = q, X_i = x]. \tag{1.2}$$

Since the treatment is binary, $p(q,x)$ equals the conditional treatment probability, $\mathbb{P}(T_i = 1 \mid Q_i = q, X_i = x)$, also known as the propensity score. When viewed as a random variable, I will denote it by $P_i = p(Q_i, X_i)$. Similarly, $r(q,x)$ denotes the conditional expectation of the outcome, and I denote it by $R_i = r(Q_i, X_i)$ when viewed as a random variable. Without further assumptions, these regression function are not directly informative about the objects of interest—the treatment effects. They are therefore known as the reduced form equations.

In general, both reduced form equations will be non-linear. The linear IV estimators that I will consider are based on a linear approximation to the true non-linear reduced form

$$R_i^L = \mathbb{E}^*[Y_i \mid Z_i, W_i] = Z_i'\pi_1 + W_i'\psi_1, \tag{1.3}$$

$$P_i^L = \mathbb{E}^*[T_i \mid Z_i, W_i] = Z_i'\pi_2 + W_i'\psi_2, \tag{1.4}$$

where $\mathbb{E}^*$ denotes population (minimum mean-squared-error) linear projection[1], and $Z_i = z(Q_i, X_i)$ and $W_i = w(X_i)$ are expansions of the original instruments and covariates, with $\dim(Z_i) = K$ and $\dim(W_i) = L$. I assume that W_i spans a column of ones. The estimators that I will consider will use these constructed instruments and covariates.

For example, in Angrist and Krueger (1991), the basic instruments Q_i were three quarter of birth indicators, and the constructed instruments Z_i were obtained by interacting Q_i with

[1]In other words, the linear projection of A_i onto B_i, $\mathbb{E}^*[A_i \mid B_i] = B_i'\gamma$, minimizes $\min_\gamma \mathbb{E}[(A_i - B_i'\gamma)^2]$. If the covariance matrix of B_i is non-singular so that $\mathbb{E}[B_iB_i']$ is invertible, then the solution is uniquely given by $\gamma = \mathbb{E}[B_iB_i']^{-1}\mathbb{E}[B_iA_i]$.

year of birth and state of birth indicators. A similar specification was used in Dobbie and Fryer (2011), who study the effect of Harlem Children Zone (HCZ) charters on educational outcomes. In particular, Dobbie and Fryer (2011) construct Z_i by interacting an indicator for living within HCZ, Q_i, with cohort, so that $Z_{i,\ell} = Q_i \mathbb{1}_{X_i=\ell}$, where ℓ indexes cohorts, $\ell \in \{1, \ldots, L\}$. If we also set $W_{i\ell} = \mathbb{1}_{X_i=\ell}$, and cohort is the only covariate that we observe, then the linear approximation is exact, and $P_i = P_i^L$. With continuous instruments and covariates, we could use series expansions to construct Z_i and W_i. Of course, we can also simply set $z(Q_i, X_i) = Q_i$ and $w(X_i) = X_i$. I make the distinction between the original instruments and covariates, (Q_i, X_i), and the constructed ones, (Z_i, W_i), because it will matter for the estimands of these estimators under treatment effect heterogeneity how exactly the instruments were constructed.

I use matrix notation to help keep the definitions and results compact. I denote the n-component vector with ith element Y_i by $\mathbf{Y}$. Similarly, let $\mathbf{T}$, $\mathbf{W}$, $\mathbf{Z}$, $\mathbf{P}$, $\mathbf{P}^L$, $\mathbf{R}$ and $\mathbf{R}^L$ denote vectors and matrices with rows T_i, W_i', Z_i', P_i, P_i^L, R_i and R_i^L. For any full-rank $n \times m$ matrix $\mathbf{A}$, let $\mathbf{H_A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ denote the associated $n \times n$ projection matrix (also known as the hat matrix), and let $\mathbf{D_A}$ be an $n \times n$ diagonal matrix with $(\mathbf{H_A})_{ii}$ on the diagonal. Let $\mathbf{I}_m$ denote the $m \times m$ identity matrix, and let $\mathbf{M_A} = \mathbf{I}_n - \mathbf{H_A}$ denote the annihilator matrix. Let $\mathbf{A}_\perp = \mathbf{M_W A}$ denote the residual from the sample projection of $\mathbf{A}$ onto $\mathbf{W}$, and let $\tilde{A}_i = A_i - \mathbb{E}^*[A_i \mid W_i]$ denote the residual from the population projection of A_i onto W_i. Also, let a.s. denote almost surely (i.e. with probability one).

1.3 Classic linear IV model and estimators

The classic linear IV model is usually defined in terms of a structural equation (see, for example Wooldridge, 2002, Chapter 5)

$$Y_i = W_i'\delta + T_i\beta + \epsilon_i, \tag{1.5}$$

where the covariates W_i and the instruments Z_i are assumed to be uncorrelated with the structural error ϵ_i:

$$\mathbb{E}[\epsilon_i W_i] = 0, \qquad \mathbb{E}[\epsilon_i Z_i] = 0. \tag{1.6}$$

The second assumption is that the instruments are relevant in the sense that the coefficient π_2 in Equation (1.4) is non-zero. The parameter of interest is β, and it represents the causal effect of T_i on Y_i. Equations (1.5)–(1.6) can be compactly written as a moment condition

$$\mathbb{E}^*[Y_i - W_i'\delta - T_i\beta \mid Z_i, W_i] = 0. \tag{1.7}$$

In this section, I use the potential outcomes framework to formulate assumptions that deliver the moment condition (1.7) and that give β a direct causal interpretation as the population average treatment effect, $\mathbb{E}[\tau_i]$. This allows me to more easily link the classic IV model to the LATE framework of Imbens and Angrist (1994). Second, I define two classes of estimators of β: the class of two-step IV estimators (that includes the two-stage least squares estimator) and the class of minimum-distance estimators (that includes LIML). This classification will be more useful when considering the behaviour of the classic IV estimators under the LATE framework than the traditional division into estimators that fit into the k-class (Nagar, 1959; Theil, 1961, 1971), and estimators that do not.

1.3.1 Assumptions underlying the classic linear IV model

Interpreting β in the moment condition (1.7) as the population average treatment effect requires three assumptions that correspond to Assumptions IV, CTE and L below. First, the instruments need to be valid in the sense that they only affect potential outcomes through their effect on the treatment. Second, we need to restrict treatment effect heterogeneity. Third, we need to make some functional form assumptions.

In order to state formally what properties valid instruments should have, we need to include Q in the definition of potential outcomes. Let $Y_i(t, q)$ be the potential outcome when individual i receives treatment t and instrument q, so that the observed outcome is given by

$Y_i = Y_i(T_i, Q_i)$.

Assumption IV.

(i) (Random assignment) $\{Y_i(t,q), T_i(q)\}_{t\in\{0,1\},q\in\mathcal{Q}} \perp\!\!\!\perp Q_i \mid X_i$;

(ii) (Exclusion restriction) $\mathbb{P}(Y_i(t,q) = Y_i(t,q') \mid X_i) = 1$ for all (t,q,q') a.s.; and

(iii) (Relevance) The distribution of P_i^L conditional on X_i is non-degenerate with positive probability.

Part (i) requires that conditional on covariates, the instruments are as good as randomly assigned in the sense that they are independent of potential outcomes and potential treatments. Part (ii) requires that the instruments only affect outcomes through their effect on the treatment. This assumption justifies writing the potential outcomes as functions of the treatment only, so that $Y_i(t) = Y_i(t,q)$. Finally, Part (iii) is a rank condition—requires that the constructed instruments Z_i have a non-zero effect on the treatment, at least for some values of covariates; it ensures that the coefficient π_2 in Equation (1.4) is non-zero. Since $Z_i = z(Q_i, X_i)$, a necessary condition is that the original instruments Q_i have a non-zero effect on the treatment.

Assumption CTE (Constant Average Treatment Effects). For all $(t,q,x) \in \{0,1\} \times \mathcal{Q} \times \mathcal{X}$, $\mathbb{E}[\tau_i \mid Q_i = q, T_i = t, X_i = x] = \beta$.

Although Assumption CTE allows the individual treatment effects τ_i to vary, it requires that the source of heterogeneity in the individual treatment effects be unrelated to observables. In particular, it does not allow individuals' treatment status to be correlated with gains from treatment, ruling out what Heckman, Urzua and Vytlacil (2006) call essential heterogeneity, or sorting on gains from treatment. This makes it implausible in many empirical applications—I will relax it in the next section when I introduce the LATE framework of Imbens and Angrist (1994).

By the Law of iterated expectations, the parameter β corresponds to the average treatment effect $\mathbb{E}[\tau_i]$. Some textbook discussions of the classic linear IV model (Wooldridge, 2002; Angrist and Pischke, 2009) use a stronger version of this assumption by imposing $\tau_i = \beta$ for

all i, ruling out *any* heterogeneity in the treatment effect, but such restrictive assumption is not needed.

Assumption CTE implies that

$$\begin{aligned} 0 &= \mathbb{E}[Y_i(1) - Y_i(0) - \beta \mid Q_i = q, T_i = t, X_i = x] \\ &= \mathbb{E}[T_i(Y_i(1) - Y_i(0) - \beta) \mid Q_i = q, T_i = t, X_i = x] \\ &= \mathbb{E}[Y_i - Y_i(0) - T_i\beta \mid Q_i = q, X_i = x], \end{aligned} \tag{1.8}$$

where the last line follows from $Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i)$ and the Law of iterated expectations. To turn Equation (1.8) into the moment condition (1.7), we need that

$$\mathbb{E}^*[Y_i(0) \mid Z_i, W_i] = \mathbb{E}^*[Y_i(0) \mid W_i]. \tag{1.9}$$

If there are no covariates beyond the intercept, so that $W_i = 1$, then this equality holds automatically. However, since Assumption IV allows for cases in which the assignment of instrument is only random conditional on covariates, it only implies that $\mathbb{E}[Y_i(0) \mid Z_i, X_i] = \mathbb{E}[Y_i(0) \mid X_i]$. If the conditional expectation $\mathbb{E}[Y_i(0) \mid X_i]$ is not linear in W_i, then controlling for W_i in a linear way does not fully control for the effect of the covariates on $Y_i(0)$. Consequently, $\tilde{Z}_i = Z_i - \mathbb{E}^*[Z_i \mid W_i]$ (part of Z_i orthogonal to W_i) may be correlated with $Y_i(0) - \mathbb{E}^*[Y_i(0) \mid W_i]$, and the coefficient on Z_i on the left-hand side of (1.9) may be non-zero. Therefore, some textbook discussions (Wooldridge, 2002; Angrist and Pischke, 2009) make the assumption that $\mathbb{E}[Y_i(0) \mid X_i] = W_i'\delta$, so that controlling for W_i in a linear way controls fully for the effect of the covariates on $Y_i(0)$. Unfortunately, this assumption has the undesirable implication that, in principle, sufficient variation in the covariates alone is enough to identify β since non-linear functions of W_i, such as squares of W_i, are valid instruments. Moreover, since it involves potential, rather than observed outcomes, it is not directly testable.

Here I focus on the other way we can ensure that $\tilde{Z}_i$ is not correlated with $Y_i(0) - \mathbb{E}^*[Y_i(0) \mid W_i]$—by restricting the expectation of Z_i conditional on X_i to be linear in W_i:[2]

[2]By the residual regression formula (1.9) holds iff $\mathbb{E}[Y_i(0)\tilde{Z}_i] = 0$. Assumption L implies that $\mathbb{E}[\tilde{Z}_i \mid X_i] = 0$.

Assumption L (Linearity). $\mathbb{E}[Z_i \mid X_i] = \mathbb{E}^*[Z_i \mid W_i]$.

Assumption L ensures that controlling for the effect of covariates on the instruments by a linear projection on W_i is as good as conditioning on X_i. There are three important special cases in which Assumption L holds automatically. First, if there are no covariates. Second, if X_i is discrete and W_i is saturated, consisting of dummies for different values of X_i. Third, if Z_i is a function of Q_i only, and Q_i is independent of X_i, in which case $\mathbb{E}[Z_i \mid X_i] = \mathbb{E}[Z_i]$. This happens, for example, when Q_i is some randomly assigned encouragement to take the treatment, and the covariates are added after the randomization to increase precision of inference.

Abadie (2003) shows that another consequence of Assumption L is that the parameter δ in Equation (1.7) can now be interpreted as providing the best linear approximation to $\mathbb{E}[Y_i(0) \mid X_i]$ in the sense of minimizing the mean-square error $\mathbb{E}[(\mathbb{E}[Y_i(0) \mid X_i] - W_i'\delta)^2]$.

1.3.2 Two-step IV estimators

An implication of the moment condition (1.7) is that β can be identified using a single instrument $\tilde{P}_i^L = \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i] = \tilde{Z}_i'\pi_2$, the linear approximation to the propensity score (1.4) with the covariates partialled out. $\tilde{P}_i^L$ can be thought of as an approximation to $\mathbb{E}[T_i \mid Q_i, X_i] - \mathbb{E}[T_i \mid X_i] = P_i - \mathbb{E}[P_i \mid X_i]$, which measures how strong the instrument assigned to individual i is (in terms of how likely it is to induce an individual into taking the treatment), relative to other instruments they could have been assigned, holding the covariates fixed. Since $\tilde{P}_i^L$ is linear in Z_i and W_i, the moment condition implies that

$$0 = \mathbb{E}^*[Y_i - W_i'\delta - T_i\beta \mid \tilde{P}_i^L] = \mathbb{E}^*[Y_i - T_i\beta \mid \tilde{P}_i^L],$$

Therefore, by the law of iterated expectations, we have $\mathbb{E}[Y_i(0)\tilde{Z}_i \mid X_i] = \mathbb{E}[\mathbb{E}[Y_i(0) \mid X_i, Z_i]\mathbb{E}[\tilde{Z}_i \mid X_i, Z_i] \mid X_i] = \mathbb{E}[Y_i(0) \mid X_i]\mathbb{E}[\tilde{Z}_i \mid X_i] = 0$ where the second equality follows from Assumption IV.

where the second equality follows from $\mathbb{E}[W_i \tilde{P}_i^L] = 0$. Rearranging this expression, we obtain

$$\beta = \frac{\mathbb{E}[\tilde{P}_i^L Y_i]}{\mathbb{E}[\tilde{P}_i^L T_i]},$$

so that the IV estimator that uses $\tilde{P}_i^L$ as a single instrument, $\hat{\beta}_{\text{IV}} = \sum_i \tilde{P}_i^L Y_i / \sum_i \tilde{P}_i^L T_i$, is consistent for β. Moreover, if the error $\epsilon_i = Y_i - W_i'\delta - T_i\beta$ is homoscedastic, so that $\text{var}(\epsilon_i^2 \mid X_i, Q_i) = \sigma^2$, then this estimator is asymptotically efficient.

Since $\tilde{P}_i^L$ is not directly observed, such an estimator is not feasible. Two step IV estimators implement a feasible version of $\hat{\beta}_{\text{IV}}$. In the first-step, they construct an estimate $\hat{P}_i$ of $\tilde{P}_i^L$. In the second step, an IV estimator that uses this constructed instrument as a single instrument is used to estimate the treatment effect:

$$\hat{\beta}_{\hat{\mathbf{P}}} = \frac{\hat{\mathbf{P}}'\mathbf{Y}}{\hat{\mathbf{P}}'\mathbf{T}}. \tag{1.10}$$

The class of two-step IV estimators is given by all estimators that admit this representation, where $\hat{\mathbf{P}}$ is a function of $\mathbf{T}$, $\mathbf{W}$ and $\mathbf{Z}$, including:

- The two-stage least squares (TSLS) estimator, which replaces π_2 and ψ_2 in (1.4) by their least-squares estimates, leading to $\hat{\mathbf{P}} = \mathbf{H}_{\mathbf{Z}_\perp}\mathbf{T}$;
- The bias-corrected two-stage least squares estimator (Nagar, 1959), which adjusts the TSLS propensity score estimator to $\hat{\mathbf{P}} = ((1-k)\mathbf{M}_{\mathbf{W}} + k\mathbf{H}_{\mathbf{Z}_\perp})\mathbf{T}$ to improve its finite-sample properties, where $k = 1/(1-(K-2)/n)$;
- The jackknife instrumental variables estimator (Phillips and Hale, 1977; Angrist *et al.*, 1999), with $\hat{\mathbf{P}} = \mathbf{M}_{\mathbf{W}}\big(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z},\mathbf{W})})^{-1}\mathbf{M}_{(\mathbf{Z},\mathbf{W})}\big)\mathbf{T}$.

Under regularity conditions, the estimation error in the first step does not matter, and all of these estimators are consistent for β, and asymptotically efficient under homoscedasticity.

1.3.3 Minimum distance estimators

Another implication of the conditional moment restriction (1.7) is that if we project Y_i and T_i onto Z_i and W_i, the coefficients on Z_i will be proportional to each other. To see this, by linearity of linear projections, we obtain:

$$\mathbb{E}^*[Y_i \mid Z_i, W_i] = W_i'\delta + \mathbb{E}^*[T_i \mid Z_i, W_i]\beta. \tag{1.11}$$

Therefore, the coefficients in the linear projections (1.3)-(1.4) are related to the coefficients (β, δ) by $\delta = \psi_1 - \psi_2\beta$, and

$$\pi_1 = \pi_2\beta. \tag{1.12}$$

This proportionality restriction can be imposed directly in estimation of β by using a minimum distance objective function

$$(\text{vec}(\hat{\Pi}) - a \otimes \pi_2)'\hat{\Phi}(\text{vec}(\hat{\Pi}) - a \otimes \pi_2), \qquad a = \begin{pmatrix} \beta \\ 1 \end{pmatrix}, \tag{1.13}$$

where $\hat{\Pi} = (\mathbf{Z}_\perp'\mathbf{Z}_\perp)^{-1}\mathbf{Z}_\perp'(\mathbf{Y}, \mathbf{T})$ is an unrestricted least-squares estimator of $\Pi = (\pi_1, \pi_2)$, and $\hat{\Phi}$ is some weight matrix. Goldberger and Olkin (1971) show that the limited information maximum likelihood (LIML) estimator minimizes this objective function if the weight matrix is given by

$$\hat{\Phi} = \hat{\Omega}^{-1} \otimes \mathbf{Z}_\perp'\mathbf{Z}_\perp/n, \qquad \hat{\Omega} = \begin{pmatrix} \mathbf{Y} & \mathbf{T} \end{pmatrix}' \mathbf{M}_{(\mathbf{Z},\mathbf{W})} \begin{pmatrix} \mathbf{Y} & \mathbf{T} \end{pmatrix} / (n - K - L).$$

Here $\hat{\Omega}$ an estimator of the covariance matrix of the reduced-form errors $V_{1i} = Y_i - \mathbb{E}^*[Y_i \mid Z_i, W_i]$ and $V_{2i} = T_i - \mathbb{E}^*[T_i \mid Z_i, W_i]$ based on the unrestricted least-squares residuals. To understand the sensitivity of minimum distance estimators to departures from the assumption of constant treatment effects (Assumption CTE), it is helpful to work with a slightly different minimum distance objective function. Define a 2-by-2 matrix

$$\Xi = \Pi'\mathbb{E}[\tilde{Z}_i\tilde{Z}_i']\Pi. \tag{1.14}$$

In Section 1.5, I will show that this matrix plays a key role in understanding the behaviour of classic linear IV estimators under the LATE framework. The proportionality restriction (1.12) implies a rank restriction on Ξ, namely that $\Xi = \Lambda aa'$, where $\Lambda = \Xi_{22} = \pi_2' \mathbb{E}[\tilde{Z}_i \tilde{Z}_i'] \pi_2$. This restriction is essentially a restriction on the second moments of $\hat{\Pi}$ if $\mathbb{E}[\tilde{Z}_i \tilde{Z}_i']$ is proportional to the identity matrix. If the weight matrix $\hat{\Phi}$ has a Kronecker structure, $\hat{\Phi} = \hat{S}^{-1} \otimes \mathbf{Z}_\perp' \mathbf{Z}_\perp / n$ for some positive definite matrix $\hat{S} \in \mathbb{R}^{2\times 2}$, minimizing the objective function (1.13) yields the same estimator of β as a minimum distance estimator based on the rank-restriction on Ξ given by[3]

$$\hat{\mathcal{D}}(\beta, \Lambda) = \operatorname{vec}(\hat{\Xi} - \Lambda aa')'(\hat{S}^{-1} \otimes \hat{S}^{-1}) \operatorname{vec}(\hat{\Xi} - \Lambda aa'), \tag{1.15}$$

where $\hat{\Xi} = (\mathbf{Y}, \mathbf{T})' \mathbf{H}_{\mathbf{Z}_\perp} (\mathbf{Y}, \mathbf{T}) / n = \hat{\Pi}' (\mathbf{Z}_\perp' \mathbf{Z}_\perp / n) \hat{\Pi}$ is an unrestricted estimator of Ξ. The class of minimum distance estimators is given by all estimators that minimize the objective function (1.15) for some unrestricted estimator $\hat{\Xi}$ of Ξ and some weight matrix $\hat{S}^{-1} \otimes \hat{S}^{-1}$. These estimators can be written as

$$\hat{\beta}_{\hat{\Xi}, \hat{S}} = \frac{\hat{\Xi}_{12} - \hat{S}_{12} \min \operatorname{eig}(\hat{S}^{-1} \hat{\Xi})}{\hat{\Xi}_{22} - \hat{S}_{22} \min \operatorname{eig}(\hat{S}^{-1} \hat{\Xi})}. \tag{1.16}$$

Apart from LIML, the class of minimum distance estimators includes:

- Ω-class estimators of Keller (1975), which, like LIML, set $\hat{\Xi} = (\mathbf{Y}, \mathbf{T})' \mathbf{H}_{\mathbf{Z}_\perp} (\mathbf{Y}, \mathbf{T}) / n$, but S is free to be any positive definite matrix. The choice $\hat{S} = \mathbf{I}_2$ leads to the symmetrically normalized two-stage least squares estimator studied in Keller (1975), Hillier (1990) and Alonso-Borrego and Arellano (1999).

- The symmetric jackknife estimator of (Bekker and Crudu, 2012), which sets

$$\hat{\Xi} = \frac{1}{n} \begin{pmatrix} \mathbf{Y} & \mathbf{T} \end{pmatrix}' \left(\mathbf{H}_{\mathbf{Z}_\perp} - \frac{1}{2} \left(\mathbf{H}_{\mathbf{Z}_\perp} \mathbf{C} + \mathbf{C}' \mathbf{H}_{\mathbf{Z}_\perp} \right) - \frac{1}{4} \mathbf{C}' \mathbf{H}_{\mathbf{W}} \mathbf{C} \right) \begin{pmatrix} \mathbf{Y} & \mathbf{T} \end{pmatrix},$$
$$\hat{S} = \frac{1}{n - K - L} \begin{pmatrix} \mathbf{Y} & \mathbf{T} \end{pmatrix}' \mathbf{M}_{(\mathbf{Z}, \mathbf{W})} \mathbf{D}_{(\mathbf{Z}, \mathbf{W})} (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}, \mathbf{W})})^{-1} \mathbf{M}_{(\mathbf{Z}, \mathbf{W})} \begin{pmatrix} \mathbf{Y} & \mathbf{T} \end{pmatrix},$$

  where $\mathbf{C} = \mathbf{D}_{(\mathbf{Z}, \mathbf{W})} (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}, \mathbf{W})})^{-1} \mathbf{M}_{(\mathbf{Z}, \mathbf{W})}$.

[3] See Kolesár (2012) for derivation.

- If there are no covariates W_i, then the HLIM estimator of Hausman *et al.* (2012) also admits this minimum distance representation, with

$$\hat{\Xi} = \frac{1}{n}\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}'(\mathbf{H_Z} - \mathbf{D_Z})\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}, \quad \hat{S} = \frac{1}{n-K}\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}'(\mathbf{M_Z} + \mathbf{D_Z})\begin{pmatrix}\mathbf{Y} & \mathbf{T}\end{pmatrix}.$$

Under homoscedasticity, any weight matrix S produces an asymptotically efficient estimator (Alonso-Borrego and Arellano, 1999). Consequently all of these estimators are asymptotically efficient under these conditions, and first-order asymptotically equivalent to the optimal two-step IV estimator. The reason for this is that under standard asymptotics, the estimator $\hat{\Xi}_{12}/\hat{\Xi}_{22}$ of β that does not use the information about β contained in Ξ_{11} is asymptotically equivalent to a minimum distance estimator that uses the optimal weight matrix. This is easy to check since, in fact, $\hat{\Xi}_{12}/\hat{\Xi}_{22}$ is the two-stage least squares estimator.

1.4 Local average treatment effects approach

Instead of restricting treatment effect heterogeneity, the local average treatment effects framework of Imbens and Angrist (1994) replaces Assumption CTE by a monotonicity assumption that restricts how a treatment response to changing the value of the instrument may vary *across* people:

Assumption M (Monotonicity). For all $q, q' \in \mathcal{Q}$ either $\mathbb{P}(T_i(q) \geq T_i(q') \mid X_i) = 1$ or $\mathbb{P}(T_i(q) \leq T_i(q') \mid X_i) = 1$ a.s.

This assumption maintains that changing the instruments from q to q' affects all individuals with the same value of X_i in the same direction—it rules out situations in which, in response to a change in Q_i, some people drop out of treatment and others select into it. If Q_i is an encouragement to take the treatment, for example, then monotonicity requires that encouraging people to take the treatment makes everyone more likely to take it. Vytlacil (2002) shows that Assumption M is equivalent to assuming a latent index model as first proposed by Heckman (1976), in which selection into the treatment is modeled by a latent

index crossing a threshold.[4]

For each value $x \in \mathcal{X}$ and for each pair (q, q'), define a local average treatment effect (LATE):

$$\tau(q, q'; x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i(q) \neq T_i(q'), X_i = x]. \tag{1.17}$$

This is the treatment effect averaged over individuals with $X_i = x$ who change their treatment status if we change their instrument from q to q'. Angrist, Imbens and Rubin (1996) refer to this set of individuals as compliers. Imbens and Angrist (1994) show that under Assumptions IV, and M, so long as $\mathbb{P}(T_i(q) \neq T_i(q') \mid X_i = x) > 0$, these local average treatment effects can be identified from the reduced form regressions:

$$\tau(q, q'; x) = \frac{r(q, x) - r(q', x)}{p(q, x) - p(q', x)}. \tag{1.18}$$

If $\mathbb{P}(T_i(q) \neq T_i(q') \mid X_i = x) = 0$, then the set of compliers that the local average treatment effect (1.17) conditions on is empty, $p(q, x) = p(q', x)$, and $\tau(q, q'; x)$ is not identified. Since Assumption IV (iii) implies that the distribution of P_i conditional on X_i is non-degenerate with positive probability, it ensures that at least some local average treatment effects are identified. On the other hand, the population average treatment effect $\mathbb{E}[\tau_i]$ is no longer identified once Assumption CTE is dropped unless the instrument Q_i is sufficiently strong to change everyone's treatment status (known as "identification at infinity"). The reason is that without restricting treatment effect heterogeneity, we have no way of computing the treatment effect for individuals who don't change their treatment status in response to a change in Q_i.

To facilitate expressing estimands or estimators based on the linear IV model in terms of local average treatment effects, it will be useful to write $\tau(q, q'; x)$ in terms of functions of the propensity score. Because of the equivalence between monotonicity and single index models, the instruments Q_i enter the model only through the propensity score (Heckman and Vytlacil, 1999; Heckman *et al.*, 2006). Therefore, $r(q, x) = \mathbb{E}[Y_i \mid P_i = p(q, x), X_i = x]$. Let $\mathcal{P}_x$

[4] In the Heckman (1976) model, the index is given by $T_i^* = p(Q_i, X_i) - U_i$, where U_i is an unobserved random variable, distributed independently of (Q_i, X_i). T_i^* is interpreted as the expected net utility of selecting into treatment, so that $T_i = 1$ if $T_i^* \geq 0$.

be the support of P_i conditional on $X_i = x$. Suppose that Q_i is discrete, so that $\mathcal{P}_x$ has finitely many support points. Let J_x be the number of support points, with $\mathcal{P}_x = \{p_{1,x} < \ldots < p_{J_x,x}\}$. Define a *marginal* local average treatment effect:

$$\alpha(p_{j,x}; x) = \frac{\mathbb{E}[Y_i \mid P_i = p_{j+1,x}, X_i = x] - \mathbb{E}[Y_i \mid P_i = p_{j,x}, X_i = x]}{p_{j+1,x} - p_{j,x}}, \qquad j = 1, \ldots, J_x - 1. \tag{1.19}$$

$\alpha(p_{j,x}; x)$ is the is the local average treatment effect for individuals who get treated when the instrument they receive corresponds to propensity score with rank higher than j but not otherwise. We can express every local average treatment effect (1.17) for which the set of compliers is non-empty in terms of these marginal LATEs. In particular, let $p(q, x) = p_{j,x}$ and that $p(q', x) = p_{j',x}$, and suppose that $j > j'$. Then we obtain:

$$\begin{aligned} \tau(q, q'; x) &= \frac{\sum_{m=j'}^{j-1} \left(\mathbb{E}[Y_i \mid P_i = p_{m+1,x}, X_i = x] - \mathbb{E}[Y_i \mid P_i = p_{m,x}, X_i = x]\right)}{p_{j,x} - p_{j',x}} \\ &= \sum_{m=j'}^{j-1} \frac{p_{m+1,x} - p_{m,x}}{p_{j,x} - p_{j',x}} \alpha(p_m; x). \end{aligned} \tag{1.20}$$

If $j' = j - 1$, then $\tau(q, q'; x) = \alpha(p_{j'}; x)$.

If the support of $\mathcal{P}_x$ is continuous, with $\mathcal{P}_x = [\underline{p}_x, \bar{p}_x]$, a similar result obtains if we replace the marginal LATEs $\alpha(p_{j,x}; x)$ by their limit as $q \to q'$, the marginal treatment effect (Heckman, 1997):

$$\tau(q, q'; x) = \frac{1}{p(q, x) - p(q', x)} \int_{p(q',x)}^{p(q,x)} \text{mte}(p; x)\, \mathrm{d}p, \quad \text{mte}(p; x) = \frac{\partial}{\partial p} \mathbb{E}[Y_i \mid P_i = p, X_i = x],$$

where the equality follows from Equation (1.18) and the fundamental theorem of calculus. To keep the exposition simple, I will focus on the case with discrete instruments and finite support $\mathcal{P}_x$. The results in this chapter generalize easily to the continuous case by replacing $\alpha(p_m; x)$ with the marginal treatment effect, $(p_{m+1,x} - p_{m,x})$ with $\mathrm{d}p$, and replacing sums with integrals.

1.5 Estimands under the LATE framework

This section presents the first main result of the chapter: the estimands of two-step IV estimators and minimum distance estimators when we do not restrict treatment effect heterogeneity.

I derive this result in two steps. First, in Lemma 1.1 below, I express their probability limits in terms of the reduced-form parameter Ξ, defined in Equation (1.14)—this result does not require any modelling assumptions. Second, I assume the local average treatment effects framework, and I express these reduced-form limits in terms of local average treatment effects.

Lemma 1.1. *Suppose that the data $\{Y_i, T_i, Q_i, Z_i, X_i, W_i\}_{i=1}^n$ are i.i.d with finite second moments.*

(i) *Consider a two-step* IV *estimator $\hat{\beta}_{\hat{\mathbf{P}}}$ that satisfies $\hat{\mathbf{P}}'\mathbf{Y}/n \xrightarrow{p} \mathbb{E}[\tilde{P}_i^L Y_i]$ and $\hat{\mathbf{P}}'\mathbf{T}/n \xrightarrow{p} \mathbb{E}[\tilde{P}_i^L T_i] \neq 0$, where $\tilde{P}_i^L = \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i]$. Then:*

$$\hat{\beta}_{\hat{\mathbf{P}}} \xrightarrow{p} \frac{\mathbb{E}[\tilde{P}_i^L Y_i]}{\mathbb{E}[\tilde{P}_i^L T_i]} = \frac{\Xi_{12}}{\Xi_{22}}.$$

(ii) *Consider the reverse two-stage least squares estimator given by $\hat{\beta}_{\text{RTSLS}} = \frac{\mathbf{Y}'\mathbf{H}_{\mathbf{Z}_\perp}\mathbf{Y}}{\mathbf{Y}'\mathbf{H}_{\mathbf{Z}_\perp}\mathbf{T}}$. Suppose that $\Xi_{12} \neq 0$ and that $\mathbb{E}[(Z_i, W_i)(Z_i, W_i)']$ is full rank. Then:*

$$\hat{\beta}_{\text{RTSLS}} \xrightarrow{p} \frac{\mathbb{E}[\tilde{R}_i^L Y_i]}{\mathbb{E}[\tilde{R}_i^L T_i]} = \frac{\Xi_{11}}{\Xi_{12}},$$

where $\tilde{R}_i^L = \mathbb{E}^*[Y_i \mid Z_i, W_i] - \mathbb{E}^*[Y_i \mid W_i]$.

(iii) *Consider a minimum distance estimator $\hat{\beta}_{\hat{\Xi},\hat{S}}$ that satisfies $\hat{S} \xrightarrow{p} S$ for some positive definite matrix S, and $\hat{\Xi} \xrightarrow{p} \Xi$. Suppose that $\Xi_{22} \neq \min \text{eig}(S^{-1}\Xi) S_{22}$. Then $\hat{\beta}_{\hat{\Xi},\hat{S}} \xrightarrow{p} \beta_S$, where β_S minimizes the objective function*

$$\mathcal{D}_S(\beta, \Lambda) = \text{vec}(\Xi - aa'\Lambda)'(S^{-1} \otimes S^{-1}) \,\text{vec}(\Xi - aa'\Lambda), \tag{1.21}$$

and it is given by

$$\beta_S = \frac{\Xi_{12} - S_{12} \min \text{eig}(S^{-1}\Xi)}{\Xi_{22} - S_{22} \min \text{eig}(S^{-1}\Xi)}.$$

Lemma 1.1 shows that understanding how the reduced-form parameter Ξ relates to local average treatment effects is the key to understanding the properties of estimators based on the classic linear IV model.

In particular, Part (i) shows that the probability limit of TSLS and other two-step IV estimators is simply given by Ξ_{12}/Ξ_{22}, the estimand of an IV estimator that uses the linear predictor of the treatment (with the effect of the covariates partialled out), $\tilde{P}_i^L$, as a single instrument. It makes a high-level assumption that the first-step estimator $\hat{P}_i$ converges to its population target, $\tilde{P}_i^L$. The primitive conditions for this depend on the estimator, but for TSLS, a sufficient condition is that $\mathbb{E}[(Z_i, W_i)(Z_i, W_i)']$ is full rank.

Part (ii) introduces a new estimator, the reverse two-stage least squares estimator (RTSLS). It is obtained as the reciprocal of the TSLS estimator in the instrumental variables model that swaps the treatment and the outcome:

$$\hat{\beta}_{\text{RTSLS}} = \left(\frac{\hat{\mathbf{R}}'_{\text{RTSLS}}\mathbf{T}}{\hat{\mathbf{R}}'_{\text{RTSLS}}\mathbf{Y}}\right)^{-1} = \frac{\mathbf{Y}'\mathbf{H}'_{\mathbf{Z}_\perp}\mathbf{Y}}{\mathbf{Y}'\mathbf{H}_{\mathbf{Z}_\perp}\mathbf{T}},$$

where $\mathbf{R}_{\text{RTSLS}} = \mathbf{H}'_{\mathbf{Z}_\perp}\mathbf{Y}$ is the first-step estimator of $\tilde{R}_i^L = R_i - \mathbb{E}^*[R_i \mid W_i]$ based on a least-squares estimation of Equation (1.3). Lemma 1.1 shows that the probability limit of this estimator is given by Ξ_{11}/Ξ_{12}, the estimand of an IV estimator that uses the linear predictor of the outcome (again with the covariates partialled out), $\tilde{R}_i^L$, as a single instrument. The reason for introducing this estimator is that one way of thinking about what a minimum distance estimand tries to do is to think of it as trying to be close to both Ξ_{12}/Ξ_{22} and Ξ_{11}/Ξ_{12}, using the weight matrix S as a distance metric.

Part (iii) formalizes this notion. The regularity condition $\Xi_{22} \neq \min \text{eig}(S^{-1}\Xi)S_{22}$ ensures that the limiting objective function has a well-defined minimum. Again, the primitive conditions for $\hat{\Xi} \xrightarrow{p} \Xi$ depend on the estimator, but for LIML, a sufficient condition is that $\mathbb{E}[(Z_i, W_i)(Z_i, W_i)']$ is full rank.

The rationale for trying to equate the two-step IV estimand Ξ_{12}/Ξ_{22} with the RTSLS estimand Ξ_{11}/Ξ_{12} is that the classic linear IV model is symmetric in Y and T; instead of instrumenting for T in Equation (1.7) like TSLS does, we can multiply it by $1/\beta$, instrument

for Y, and take the reciprocal of the resulting estimator, obtaining RTSLS. Both TSLS and RTSLS converge to the same probability limit, equal to the population average treatment effect, so that $\Xi_{11}/\Xi_{12} = \Xi_{12}/\Xi_{22} = \beta$. As a result, Ξ is reduced rank, and there are no trade-offs in how close we can be to Ξ_{12}/Ξ_{22} and Ξ_{11}/Ξ_{12}; the weight matrix S does not matter, $\min \operatorname{eig}(S^{-1}\Xi) = 0$ for any positive-definite weight matrix S and all minimum distance estimators converge to the population average treatment effect β. By pooling the information about β contained in TSLS with the information contained in RTSLS, minimum distance estimators have more attractive finite sample properties in classic IV model than two-step IV estimators, which don't use information about β contained in RTSLS (Phillips, 1983; Hillier, 1990). They are also more efficient under many instrument asymptotics (Hausman *et al.*, 2012).

The key question is how the interpretation of two-step IV, RTSLS, and minimum distance estimands changes under the LATE framework when Assumption CTE in the classic IV model is replaced by the Assumption M. I first answer this question for two-step IV and RTSLS estimands in Theorem 1.1 and Corollary 1.1 below by expressing the two ratios Ξ_{11}/Ξ_{12} and Ξ_{12}/Ξ_{22} in terms of the marginal local average treatment effects $\alpha(\cdot)$ defined in Equation (1.19).

Theorem 1.1. *Suppose that Assumptions IV, M and L hold. Let F^X denote the distribution of X_i. Then*

$$\frac{\Xi_{12}}{\Xi_{22}} = \int \sum_{j=1}^{J_x-1} \frac{\theta_j(x)}{\int \sum_{j=1}^{J_x-1} \theta_j(x) \,\mathrm{d}F^X(x)} \alpha(p_{j,x}; x) \,\mathrm{d}F^X(x),$$

and, if $\Xi_{12} \neq 0$

$$\frac{\Xi_{11}}{\Xi_{12}} = \int \sum_{j=1}^{J_x-1} \frac{\zeta_j(x)}{\int \sum_{j=1}^{J_x-1} \zeta_j(x) \,\mathrm{d}F^X(x)} \alpha(p_{j,x}; x) \,\mathrm{d}F^X(x),$$

where

$$\theta_j(x) = (p_{j+1,x} - p_{j,x}) \mathbb{P}(P_i > p_{j,x} \mid X_i = x) \mathbb{E}[\tilde{P}_i^L \mid X_i = x, P_i > p_{j,x}],$$
$$\zeta_j(x) = (p_{j+1,x} - p_{j,x}) \mathbb{P}(P_i > p_{j,x} \mid X_i = x) \mathbb{E}[\tilde{R}_i^L \mid X_i = x, P_i > p_{j,x}].$$

Theorem 1.1 shows that both Ξ_{12}/Ξ_{22} and Ξ_{11}/Ξ_{12} can be expressed as an affine combination of (marginal) local average treatment effects (the weights integrate to one, but are not necessarily positive).

For the two-step IV weights $\theta_j(x)/\int\sum_{j=1}^{J_x-1}\theta_j(x)\,\mathrm{d}F^X(x)$ to be positive, we need that the single instrument $\tilde{P}_i^L$ is monotone in the propensity score P_i. This ensures that the last term $\mathbb{E}[\tilde{P}_i^L \mid X_i = x, P_i > p_{j,x}]$ is always positive. In other words, we need the linear approximation P_i^L to the true propensity score P_i to be good enough in the sense that changing the value of $\tilde{P}_i^L$ does not induce two-way flows in and out of treatment (see Heckman and Vytlacil (2005) and Heckman *et al.* (2006) for discussion of this issue). If the linear approximation to the propensity score is exact, so that $P_i = P_i^L$, then the weights are guaranteed to be positive. A leading case in which this condition holds automatically is when Q_i and X_i are both finite, and we estimate a saturated model in which the instruments Z_i are generated by interacting indicators for different values of Q_i with indicators for different values of X_i. In this case, Angrist and Imbens (1995) obtain a similar expression for the weights $\theta_j(x)/\int\sum_{j=1}^{J_x-1}\theta_j(x)\,\mathrm{d}F^X(x)$.

Similarly, the RTSLS weights $\zeta_j(x)/\int\sum_{j=1}^{J_x-1}\zeta_j(x)\,\mathrm{d}F^X(x)$ are positive if the single instrument $\tilde{R}_i^L = \mathbb{E}^*[Y_i \mid W_i, Z_i] - \mathbb{E}^*[Y_i \mid W_i]$ used by RTSLS is monotone in the propensity score P_i. The next corollary gives a necessary and sufficient condition for this condition to hold if the linear approximations (1.3)–(1.4) are exact.

Corollary 1.1. *Suppose that the linear approximations* (1.3)–(1.4) *are exact, so that* $\mathbb{E}[Y_i \mid Q_i, X_i] = \mathbb{E}^*[Y_i \mid Z_i, W_i]$ *and* $\mathbb{E}[T_i \mid Q_i, X_i] = \mathbb{E}^*[T_i \mid Z_i, W_i]$*, and that Assumptions IV, M, and L hold. Then the weights* $\theta_j(x)/\int\sum_{j=1}^{J_x-1}\theta_j(x)\,\mathrm{d}F^X(x)$ *are positive, and the weights* $\zeta_j(x)/\int\sum_{j=1}^{J_x-1}\zeta_j(x)\,\mathrm{d}F^X(x)$ *are positive if all marginal* LATES $\{\alpha(p_j(x); x)\}$ *have the same sign. In the special case that* $J_x = 2$ *for all* x*,*

$$\theta_1(x) = \mathrm{var}(P_i \mid X_i = x), \qquad \zeta_1(x) = \mathrm{var}(P_i \mid X_i = x)\alpha(p_{1,x}; x).$$

The proof relies on the fact that if the linear approximations (1.3)–(1.4) are exact, then the

conditional expectation of $R_i = R_i^L$ can be decomposed as

$$\mathbb{E}[R_i^L \mid P_i = p_{j,x}, X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] + \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}),$$

so that R_i^L is only monotone in the propensity score if the marginal LATEs $\alpha(\cdot)$ all have the same sign. The other implication of this decomposition is that it demonstrates that the conditional expectation of the instrument R_i^L, and hence the term $\mathbb{E}[\tilde{R}_i^L \mid X_i = x, P_i > p_{j,x}]$ depend on the size of the marginal treatment effects $\alpha(\cdot)$. In the special case that Q_i is binary, so that $J_x = 2$, and the instruments Z_i are generated by interacting Q_i with the covariates, this results in the weights ζ to be exactly equal to the product of the marginal treatment effect with the two-step IV weights θ. Therefore, larger local average treatment effects receive more weight, and negative local average treatment effects receive a negative weight in this case.

Taken together, Lemma 1.1, Theorem 1.1 and Corollary 1.1 show that under Assumptions IV, M and L, two-step IV estimators estimate a convex combination of local average treatment effects, so long as the linear approximation P_i^L to the true propensity score P_i is monotone in P_i. In the special case with a binary Q_i, Corollary 1.1 shows that these weights are given by the variance of the propensity score, so that better identified LATEs receive more weight. If in fact all LATEs are equal, then this weighting scheme ensures that under homoscedasticity, asymptotic variance of two-step IV estimators is minimized. On the other hand, the weighting used by the RTSLS estimand is different, depends on the size of the local average treatment effects, and may result in an estimand outside of the convex hull of LATEs if some LATEs are positive and some are negative.

Because it gives more weight to larger LATEs, the RTSLS estimand will always be larger than the two-step iv estimand. This result holds in general by the Cauchy-Schwarz inequality since Ξ is a covariance matrix of $(\tilde{R}_i^L, \tilde{P}_i^L)$,

$$\Xi = \mathbb{E}[(\tilde{R}_i^L, \tilde{P}_i^L)'(\tilde{R}_i^L, \tilde{P}_i^L)],$$

so that $\Xi_{11}\Xi_{22} \geq \Xi_{12}^2$. Hence, $\Xi_{11}/\Xi_{12} \geq \Xi_{12}/\Xi_{22}$, with equality only if $\tilde{P}_i^L$ is perfectly correlated with $\tilde{R}_i^L$, in which case the RTSLS weights are proportional to the two-step IV weights. There are only two ways how this can happen: either all local average treatment effects are equal, or else the dimension of Z_i is one, so that the IV model (1.7) is exactly identified. In general with more than one instrument, it will be the case that $\Xi_{11}/\Xi_{12} > \Xi_{12}/\Xi_{22}$.

The result that the RTSLS estimand is in general different from the two-step IV estimand has important implications for minimum distance estimators. On the one hand, combining RTSLS with TSLS leads to more attractive properties of minimum distance estimators in the classic IV model under which $\Xi_{11}/\Xi_{12} = \Xi_{12}/\Xi_{22}$. On the other hand, trying to equate RTSLS and TSLS when their estimands are in fact different makes minimum distance estimators unattractive under treatment effect heterogeneity; as I discuss next, it may cause the minimum distance estimands to no longer correspond to a causal effect.

If the local average treatment effects are not all equal, then $\Xi_{11}/\Xi_{12} \neq \Xi_{12}/\Xi_{22}$, and the probability limit of a minimum distance estimator depends on the weight matrix S. If the weight matrix is diagonal, then the minimum distance estimand lies between TSLS and RTSLS estimands—this was first shown in Zellner (1970) in an errors-in-variables context. Therefore, the symmetrically normalized two-stage least squares estimator (see page 19 for definition), for example, which uses the identity matrix as a weight matrix will always lie between two-step IV and RTSLS estimands. The relative weight given to the RTSLS and TSLS estimands depends on the ratio S_{11}/S_{22}. In particular if the ratio S_{11}/S_{22} is small, then the penalty from being far away from Ξ_{11}/Ξ_{12} is large, so the minimum distance estimand will be close to the RTSLS estimand. On the other hand, if S_{11}/S_{22} is large, then the minimum distance estimand will be close to the two-step IV estimand (see Zellner (1970) and Keller (1975) for a detailed discussion). Heuristically, if we concentrate Λ out of the objective function $\mathcal{D}_S(\beta, \Lambda)$ given in (1.21), we obtain that

$$\beta_S = \operatorname*{argmin}_{\beta} \frac{\Xi_{22}\beta^2 - 2\Xi_{12}\beta + \Xi_{11}}{S_{11} + S_{22}\beta^2}.$$

Now, if we set $S_{22} = 0$, then $\beta_S = \Xi_{12}/\Xi_{22}$, and if we set $S_{11} = 0$, then we obtain $\beta_S = \Xi_{11}/\Xi_{12}$.

If S is non-diagonal, however, then the minimum distance estimand may lie outside the interval formed by the two-step IV and RTSLS estimands. This is typically the case for LIML, for which S equals the covariance matrix of the reduced-form errors, which is typically non-diagonal. To see how this may happen, consider a simple model in which we observe draws of a vector A_i, distributed according to the bivariate Normal distribution with mean $(\mu_1, \mu_2)'$ and covariance matrix Ω. If $\mu_1 = \mu_2$, and Ω is known, then the optimal estimator is given by:

$$\hat{\mu} = \frac{\iota'\Omega^{-1}\overline{A}}{\iota'\Omega^{-1}\iota'}, \qquad \iota = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

where $\overline{A} = n^{-1}\sum_{i=1}^{n} A_i$. The probability limit of this estimator is given by

$$\mu_\Omega = \frac{(\Omega_{22} - \Omega_{12})\mu_1 + (\Omega_{11} - \Omega_{12})\mu_2}{\Omega_{22} + \Omega_{11} - 2\Omega_{12}}.$$

If $\Omega_{12} = 0$, then μ_Ω lies between μ_1 and μ_2. If, however, Ω is non-diagonal, then this may no longer be the case—if, for example, $\mu_2 = 0$ and μ_1 is positive, then μ_Ω will be negative if $\Omega_{22} < \Omega_{12}$.

There are two ways, therefore, in which a minimum distance estimand may end up being outside of the convex hull of LATEs. First, if some LATEs are positive and some are negative, and the RTSLS estimand is outside of the convex hull, then so long as the weight matrix S gives sufficient weight to RTSLS, the minimum distance estimand will also be outside of the convex hull. Second, even if the RTSLS estimand lies inside the convex hull, if the weight matrix S is non-diagonal, the minimum distance estimand may end up being outside of the convex hull. These possibilities make LIML and other minimum distance estimators an unattractive estimator choice in settings with possible treatment effect heterogeneity.

1.6 Estimation with many instruments

In this section, I derive the second main result of the chapter that a version of the jackknife instrumental variables estimator, the unbiased jackknife instrumental variables estimator (UJIVE), is consistent for a convex combination of LATEs under a many instrument asymptotic sequence in which both the number of instruments and the number of covariates is allowed to increase in proportion with the sample size. In settings with many instruments and treatment effect heterogeneity, UJIVE is therefore a more attractive estimator than TSLS, which is inconsistent under many instrument asymptotics. It is also more attractive than LIML, the standard alternative to TSLS when many instruments are used, since, as shown in Section 1.5, LIML may converge to a quantity outside of the convex hull of local average treatment effects even under standard asymptotics.

To illustrate the issues that arise with a large number of instruments, as well as to motivate UJIVE, I first discuss a simple example in which the instruments Z_i are indicators for group membership. I then give the general consistency theorem.

1.6.1 A simple example with groups as instruments

Consider the problem of estimating the effect of incarceration on post-release criminality, as in Aizer and Doyle, Jr. (2011) and Nagin and Snodgrass (2011). The identification strategy in these papers relies on the fact that cases are randomly assigned to judges who vary in their sentencing severity. This suggests using judge indicators as instruments for incarceration.[5] Hence, with $K+1$ judges, $\dim(Z_i) = K$, and $Z_{ik} = \mathbb{1}_{Q_i=k}$, where Q_i denotes the judge assigned to individual i (I omit the indicator for the last judge so that we can include the intercept). In this context, the monotonicity assumption requires that the judges can be ordered in terms of how strict they are. The local average treatment effects are defined for each pair of judges and correspond to the average treatment effect for individuals who would get incarcerated if assigned to the stricter judge of the two, but would not

[5] A similar strategy is also used in Dobbie and Song (2012), who study the effect of being granted bankruptcy protection on subsequent earnings, using judge indicators as instruments.

get incarcerated if assigned to the more lenient judge. If the effect of incarceration for more serious offenders (who get incarcerated unless assigned to the most lenient judges) is different from the effect for individuals who committed less serious crimes (who only get incarcerated if assigned to the strictest judges), then these LATEs will differ.

In the absence of covariates (beyond the intercept), the propensity score for individual i is given by $P_i = \mathbb{E}[T_i \mid Q_i]$, and it corresponds simply to the incarceration propensity of the judge assigned to individual i. Because the first stage is saturated, the linear approximation (1.4) is exact, and $\tilde{P}_i^L = \tilde{P}_i = P_i - \mathbb{E}[P_i]$. Intuitively, $\tilde{P}_i$ measures how strict the judge assigned to individual i is compared to other judges.

Let J_k denote the number of cases assigned to judge k. The two-stage least squares estimator of $\hat{P}_{i,\text{TSLS}} = (\mathbf{H}_{\mathbf{Z}_\perp}\mathbf{T})_i$ of $\tilde{P}_i$ can in this example be written as

$$\hat{P}_{i,\text{TSLS}} = \hat{T}_{i,\text{TSLS}} - n^{-1}\sum_{j=1}^{n}\hat{T}_{j,\text{TSLS}} = \hat{T}_{i,\text{TSLS}} - n^{-1}\sum_{j=1}^{n}T_j,$$

where $\hat{T}_{i,\text{TSLS}} = J_{Q_i}^{-1}\sum_{j:\,Q_j=Q_i}T_j$ is the sample incarceration rate for the judge assigned to individual i. It is the predictor of T_i based on least-squares estimation of the first-stage (1.4), and it is the simplest estimator of P_i. The resulting TSLS estimator is given by

$$\hat{\beta}_{\text{TSLS}} = \frac{n^{-1}\sum_i \hat{P}_{i,\text{TSLS}}Y_i}{n^{-1}\sum_i \hat{P}_{i,\text{TSLS}}T_i}. \tag{1.22}$$

There are two basic ways of doing asymptotics in this setting. The first option is to let the number of cases per judge grow to infinity while keeping the number of judges fixed. This corresponds to the standard asymptotics. As the number of cases per judge Q_i increases, $\hat{T}_{i,\text{TSLS}} \xrightarrow{p} P_i = p_{Q_i}$, and the numerator and the denominator in (1.22) converge to $\mathbb{E}[\tilde{P}_iY_i]$ and $\mathbb{E}[\tilde{P}_iT_i]$, respectively. By Lemma 1.1 and Theorem 1.1, $\hat{\beta}_{\text{TSLS}}$ therefore converges to a weighted average of local average treatment effects. However, with a large number of judges and small number of cases per judge, these asymptotics do not capture the finite-sample properties of the estimator very well.

The other possibility is to keep the number of cases per judge, J_k, fixed, and let the number of judges $K \to \infty$. This corresponds to the many instrument asymptotics (Kunitomo,

1980; Morimune, 1983; Bekker, 1994) that let the dimension of Z_i increase in proportion with the sample size. Under these asymptotics, P_i can no longer be consistently estimated, and so the exact way in which it is estimated will matter. The problem with the TSLS estimator $\hat{T}_{i,\text{TSLS}}$ is that since it includes own observation T_i, its estimation error is correlated with Y_i and T_i. As a result, the numerator and the denominator in (1.22) no longer converge to $\mathbb{E}[\tilde{P}_i Y_i]$ and $\mathbb{E}[\tilde{P}_i T_i]$. To see this, let $V_{1,i} = Y_i - R_i$ and $V_{2,i} = T_i - P_i$ denote errors in the reduced form (1.1)–(1.2), and let $K/n \to \kappa > 0$, so that the average number of cases per judge converges to $1/\kappa < \infty$. Then we can write $\hat{T}_{i,\text{TSLS}} = P_i + J_{Q_i}^{-1} \sum_{j:\, Q_j = Q_i} V_{2,j}$. We have

$$\begin{aligned}
\frac{1}{n}\sum_i \hat{P}_{i,\text{TSLS}} Y_i &= \frac{1}{n}\sum_i \hat{T}_{i,\text{TSLS}} Y_i - \frac{1}{n^2}\sum_i \sum_j T_j Y_i \\
&= \frac{K}{n}\left(\frac{1}{K}\sum_k \frac{1}{J_k} \sum_{j:\, Q_j = k} V_{2,j} \sum_{i:\, Q_i = k} Y_i\right) + \left(\frac{1}{n}\sum_i P_i Y_i\right) - \left(\frac{1}{n}\sum_j T_j\right)\left(\frac{1}{n}\sum_i Y_i\right) \\
&\xrightarrow{p} \kappa \operatorname{cov}(V_{2,i}, V_{1,i}) + \mathbb{E}[Y_i \tilde{P}_i],
\end{aligned} \tag{1.23}$$

where the last line follows from the law of large numbers applied to all four expressions in parentheses, and the fact that $\mathbb{E}[\frac{1}{J_k}\sum_{j:\, Q_j = k} V_{2,j} \sum_{i:\, Q_i = k} Y_i] = \mathbb{E}[V_{2,i} Y_i] = \mathbb{E}[V_{2,i} V_{1,i}]$. Similarly, for the denominator, $n^{-1}\sum_i \hat{P}_{i,\text{TSLS}} Y_i \xrightarrow{p} \kappa \operatorname{var}(V_{2,i}) + \mathbb{E}[T_i \tilde{P}_i]$. Therefore, TSLS is inconsistent for its target, $\mathbb{E}[\tilde{P}_i Y_i]/\mathbb{E}[\tilde{P}_i T_i]$.

There are two basic ways of adjusting the TSLS estimator to make it work under many instruments. First is to estimate the unconditional covariance matrix of $V_i = (V_{1,i}, V_{2,i})$ and subtract an estimate of the bias. This is exactly the idea behind the bias-corrected two-stage least squares estimator of Nagar (1959) and Donald and Newey (2001). Unfortunately, the estimator of the bias is only consistent under homoscedasticity (Bekker and van der Ploeg, 2005; Ackerberg and Devereux, 2009), and it is unclear how to estimate $\operatorname{var}(V_i)$ consistently when $\operatorname{var}(V_i \mid Q_i, X_i)$ is heteroscedastic. With binary T_i, $\operatorname{var}(V_{2i} \mid Q_i, X_i)$ is always heteroscedastic, so this solution is not satisfactory.

The second approach is to change the estimator of P_i so that it does not include own observation T_i. This is the idea behind the (leave-one-out) jackknife instrumental variables

estimator (JIVE, Phillips and Hale, 1977; Angrist *et al.*, 1999). It replaces $\hat{T}_{i,\text{TSLS}}$ with $\hat{T}_{i,\text{JIVE}} = (J_{Q_i} - 1)^{-1} \sum_{j:\, Q_j = Q_i, j \neq i} T_j$, the sample incarceration rate for judge Q_i with the observation on individual i excluded. The JIVE estimator of $\tilde{P}_i$ is given by

$$\hat{P}_{i,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - n^{-1} \sum_{j=1}^{n} \hat{T}_{j,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - n^{-1} \sum_{j=1}^{n} T_j. \tag{1.24}$$

The estimation error $P_i - \hat{T}_{i,\text{JIVE}} = \sum_{j:\, Q_j = Q_i, j \neq i} V_j$ is no longer correlated with T_i or Y_i, and the JIVE estimator is consistent for a convex combination of LATEs under both types of asymptotics.

So far, the discussion has abstracted from the presence of covariates. However, in practice judges are only randomly assigned at the county level. Therefore, with data from several counties, we need to include county indicators (sometimes called "fixed effects") as covariates. Hence, with L counties, $\dim(W_i) = L$, and $W_{i\ell} = \mathbb{1}_{X_i = \ell}$, where X_i denotes the county of individual i. The propensity score P_i still corresponds to the incarceration propensity of judge Q_i. However, we now have $\tilde{P}_i^L = \tilde{P}_i = P_i - \mathbb{E}[P_i \mid X_i]$, so that $\tilde{P}_i$ measures how strict judge Q_i is compared to other judges that individual i could have been assigned in the county. The JIVE estimator of P_i now becomes

$$\hat{P}_{i,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - C_{X_i}^{-1} \sum_{j:\, X_j = X_i} \hat{T}_{j,\text{JIVE}} = \hat{T}_{i,\text{JIVE}} - C_{X_i}^{-1} \sum_{j:\, X_j = X_i} T_j,$$

where C_ℓ is the number of cases in county ℓ. With a large number of counties, a natural way of thinking about the sampling is to let the number of counties $L \to \infty$, while keeping the number of judges per county and the number of cases per judge fixed. This is similar to the many instrument asymptotics in that the number of judges increases in proportion to the sample size, $K/n \to \kappa > 0$, except that instead of keeping the number of counties fixed, we also let them to grow in proportion with sample size, so that $L/n \to \lambda$. This modification of the many instrument asymptotic sequence was proposed by Anatolyev (2011) and Kolesár *et al.* (2011), and it is also used in Chetty *et al.* (2011).

Under these asymptotics, the JIVE estimator is biased. The problem is not its estimate of the propensity score—we still have that $n^{-1} \sum_i \hat{T}_{i,\text{JIVE}} Y_i \xrightarrow{p} \mathbb{E}[P_i Y_i]$, and $n^{-1} \sum_i \hat{T}_{i,\text{JIVE}} T_i \xrightarrow{p}$

$\mathbb{E}[P_i T_i]$. Instead, the source of bias comes from its estimate of the average strictness of judges in county X_i, $C_{X_i}^{-1} \sum_{j:\, X_j = X_i} T_j$. By the same logic as in the case of TSLS with many instruments, the problem is that this estimate includes own observation T_i, so that the estimation error $\mathbb{E}[P_i \mid X_i] - C_{X_i}^{-1} \sum_{j:\, X_j = X_i} T_j$ is correlated with Y_i and T_i. By arguments similar to those used to derive Equation (1.23), we have

$$\hat{\beta}_{\text{JIVE}} = \frac{n^{-1} \sum_i \hat{P}_{i,\text{JIVE}} Y_i}{n^{-1} \sum_i \hat{P}_{i,\text{JIVE}} T_i} \xrightarrow{p} \frac{\mathbb{E}[\tilde{P}_i Y_i] - \lambda \operatorname{cov}(V_{1,i} V_{2,i})}{\mathbb{E}[\tilde{P}_i T_i] - \lambda \operatorname{var}(V_{2,i})}. \tag{1.25}$$

This probability limit may differ substantially from the target $\mathbb{E}[\tilde{P}_i Y_i] / \mathbb{E}[\tilde{P}_i T_i]$, especially in settings in which the Rothenberg (1984) concentration parameter $\mathbb{E}[\tilde{P}_i T_i] / \operatorname{var}(V_{2,i}) = \pi_2' \mathbb{E}[\tilde{Z}_i \tilde{Z}_i{}'] \pi_2 / \operatorname{var}(V_{2,i})$ is small. As a result, JIVE can be severely biased in finite samples as demonstrated by Ackerberg and Devereux (2009).

The unbiased jackknife instrumental variables estimator (UJIVE) that I propose solves the bias problem of JIVE by also leaving out own observation when estimating $\mathbb{E}[P_i \mid X_i]$:

$$\hat{P}_{i,\text{UJIVE}} = \hat{T}_{i,\text{JIVE}} - \frac{1}{C_{X_i} - 1} \sum_{j:\, X_j = X_i, j \neq i} T_j.$$

Intuitively, $\hat{P}_i$ is a sample measure of how strict judge Q_i is relative to other judges in country X_i in a sample that excludes individual i. This estimator of $\hat{P}_i$ was first used in Chetty *et al.* (2011) in a setting with the same formal structure as the current example. In particular, Chetty *et al.* (2011) used classroom indicators as instruments for test score, conditioning on schools. The next subsection gives a general formula for UJIVE, and proves that it is consistent under many instrument asymptotics that also allow for many covariates.

1.6.2 Consistency of UJIVE under many instruments

Consider now the general case. Let ϕ denote the coefficient on W_i in the linear projection $\mathbb{E}^*[T_i \mid W_i]$. To define UJIVE, decompose $\tilde{P}_i^L$, the linear approximation to the propensity

score with the effect of covariates partialled out, as

$$\begin{aligned}\tilde{P}_i^L &= \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i] \\ &= Z_i'\pi_2 + W_i'\psi_2 - W_i'\phi.\end{aligned}$$

Let $\hat{\pi}_{2\backslash i}$ and $\hat{\psi}_{2\backslash i}$ be the least-squares estimates of π_2 and ψ_2 based on a sample with observation i removed. Similarly, let $\hat{\phi}_{\backslash i}$ be the least-squares estimate of ϕ based on a sample with observation i removed. The UJIVE estimator is a two-step IV estimator with the first-step estimator of $\tilde{P}_i^L$ given by

$$\hat{P}_{i,\text{UJIVE}} = Z_i'\hat{\pi}_{2\backslash i} + W_i'\hat{\psi}_{2\backslash i} - W_i'\hat{\phi}_{\backslash i}.$$

In matrix notation

$$\hat{\mathbf{P}}_{\text{UJIVE}} = \hat{\mathbf{T}}_{\text{UJIVE}} - (\mathbf{I}_n - \mathbf{D}_{\mathbf{W}})^{-1}(\mathbf{H}_{\mathbf{W}} - \mathbf{D}_{\mathbf{W}})\mathbf{T},$$

where $\hat{\mathbf{T}}_{\text{UJIVE}} = (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z},\mathbf{W})})^{-1}(\mathbf{H}_{(\mathbf{Z},\mathbf{W})} - \mathbf{D}_{(\mathbf{Z},\mathbf{W})})\mathbf{T}$. Using $\hat{\mathbf{P}}_{\text{UJIVE}}$ as a single instrument in an IV estimator then yields

$$\hat{\beta}_{\text{UJIVE}} = \frac{\hat{\mathbf{P}}_{\text{UJIVE}}'\mathbf{Y}}{\hat{\mathbf{P}}_{\text{UJIVE}}'\mathbf{T}}.$$

In contrast, while the JIVE estimator of $\mathbb{E}^*[Y_i \mid Z_i, W_i]$ is identical to $\hat{\mathbf{T}}_{\text{UJIVE}}$, its estimator of $\mathbb{E}^*[Y_i \mid W_i]$ is given by a sample projection of $\hat{\mathbf{T}}_{\text{UJIVE}}$ onto $\mathbf{W}$, so that $\hat{\mathbf{P}}_{\text{JIVE}} = \hat{\mathbf{T}}_{\text{UJIVE}} - \mathbf{H}_{\mathbf{W}}\hat{\mathbf{T}}_{\text{UJIVE}}$ (see the JIVE formula on page 17).

To formally define the many instrument asymptotic framework, I need to allow the distribution of random variables to change with the sample size. To reflect this, let the random variables be indexed by n, so that, for instance, $\mathbf{Y}_n = (Y_{n,1}, \ldots, Y_{n,n})'$ denotes the vector of observed outcomes when the sample size is n. In addition, let $P_{n,i}^X = \mathbb{E}[T_{n,i} \mid X_{n,i}]$ and $R_{n,i}^X = \mathbb{E}[Y_{n,i} \mid X_{n,i}]$ denote the expectations of $T_{n,i}$ and $Y_{n,i}$ conditional on $X_{n,i}$ only, so that $P_{n,i}^X = \mathbb{E}[P_{n,i} \mid X_{n,i}]$ and $R_{n,i}^X = \mathbb{E}[R_{n,i} \mid X_{n,i}]$. The many instrument asymptotic framework I consider is summarized by the following assumptions:

Assumption R (Regularity conditions).

(i) $\{(Y_{n,i}, T_{n,i}, X_{n,i}, Q_{n,i}) \colon i = 1, \ldots, n\}_{n \geq 1}$ is a triangular array of i.i.d. random variables, the nth row having distribution $F_n^{Y,T,X,Q}$. $F_n^{Y,T,X,Q}$ converges in distribution to $F^{Y,T,X,Q}$;

(ii) There is a positive constant C_1, such that $\sup_n \sup_{i \leq n} \operatorname{var}(Y_{n,i} \mid Q_{n,i}, X_{n,i}) \leq C_1$, and $\sup_n \sup_{i \leq n} \operatorname{var}(Y_{n,i} \mid X_{n,i}) \leq C_1$ a.s. Also, as $n \to \infty$,

$$\mathbb{E}[(R_{n,i}^2, P_{n,i}^2, |R_{n,i} P_{n,i}|)] \to \mathbb{E}[(R^2, P^2, |RP|] < \infty,$$
$$\mathbb{E}[((R_{n,i}^X)^2, (P_{n,i}^X)^2, |R_{n,i}^X P_{n,i}^X|)] \to \mathbb{E}[((R^X)^2, (P^X)^2, |R^X P^X|] < \infty,$$

where (R, P, R^X, P^X) is distributed according to the limiting distribution F^{R,P,R^X,P^X}; and

(iii) $\operatorname{rank}(\mathbf{Z}_n, \mathbf{W}_n) = K + L$ and $(\mathbf{H}_{(\mathbf{Z}_n, \mathbf{W}_n)})_{ii} < C_2$ for some $C_2 < 1$ a.s., where $Z_{n,i} = z(Q_{n,i}, X_{n,i})$ and $W_{n,i} = w(X_{n,i})$, with $\dim(Z_{n,i}) = K$ and $\dim(W_{n,i}) = L$. The functions z and w may depend on n.

Assumption MI (Many instruments). As $n \to \infty$:

(i) $K/n \to \kappa$ and $L/n \to \lambda$ for some $\kappa, \lambda \geq 0$;

(ii) $\sum_i (\mathbb{E}[T_{n,i} \mid X_{n,i}] - \mathbb{E}^*[T_{n,i} \mid W_{n,i}])^2 / n \to 0$ a.s.; and

(iii) $\sum_i (\mathbb{E}[T_{n,i} \mid Q_{n,i}, X_{n,i}] - \mathbb{E}^*[T_{n,i} \mid Z_{n,i}, W_{n,i}])^2 / n \to 0$ a.s.

Assumption R (i) allows the distribution of the data to change with the sample size, converging to some limiting distribution $F^{Y,T,X,Q}$. Part (ii) requires that the second moments of conditional expectations of $Y_{n,i}$ and $T_{n,i}$ exist and are well-behaved in the limit. It is necessary for sample averages such as $n^{-1} \sum_{i=1}^n R_{n,i}^2$ to have a well-specified probability limit. The restriction $\operatorname{rank}(\mathbf{Z}, \mathbf{W}) = K + L$ in Part (iii) is a normalization. The assumption that $(\mathbf{H}_{(\mathbf{Z},\mathbf{W})})_{ii} < C_2$ requires that no single observation has too much leverage. It implies that $(K + L)/n < C_2$ since $n^{-1} \sum_i (\mathbf{H}_{(\mathbf{Z},\mathbf{W})})_{ii} = (K + L)/n$.

Assumption MI (i) generalizes the many instrument asymptotic sequence by also allowing the number of covariates to increase with the sample size. In terms of the incarceration example, the original Bekker (1994) many instruments sequence keeps the number of counties as well as the number of cases per judge fixed, and lets the number of judges per county increase to infinity. Under Assumption MI, we can think of generating the data by

sampling L counties form some large population of counties. In Angrist and Krueger (1991), where Z_i is generated by interacting quarter of birth with L state of birth and year of birth indicators, Assumption MI (i) lets the number of states and years $L \to \infty$, while keeping the number of individuals observed in each state and year fixed. Finally, Assumption MI also accommodates models in which z and w are some approximating functions, such as splines or polynomials in the basic instruments and covariates Q_i and X_i. This corresponds to fixing the distribution of the data, so that $F_n^{Y,T,X,Q} = F^{Y,T,X,Q}$, and letting the number of terms in the approximating functions w and z increase with the sample size. Parts (ii)–(iii) then require that these approximating functions get to their population targets in the limit, and allow me to relax the requirement imposed by Assumption L that expectation of Z_i conditional on X_i is exactly linear in W_i in the sample. These conditions are similar to the assumptions in Bekker (1994) and Hansen, Hausman and Newey (2008).

Note that I do not make any assumptions about the coefficients on $Z_{n,i}$ and $W_{n,i}$ in the projections $\mathbb{E}^*[T_{n,i} \mid W_{n,i}, Z_{n,i}]$ and $\mathbb{E}^*[T_{n,i} \mid W_{n,i}]$. Under additional assumptions, such as sparsity (only few coefficients in these linear projections matter), approximations to $\tilde{P}_{n,i}^L$ other than $\hat{P}_{n,i,\text{UJIVE}}$ will work (see, for example, Belloni, Chen, Chernozhukov and Hansen, 2012).

Theorem 1.2. *Suppose that Assumptions R and MI hold, and that the limiting distribution $F^{Y,T,X,Q}$ of the data satisfies Assumptions IV, and M. Then:*

$$\begin{aligned}\hat{\beta}_{\text{UJIVE}} &\xrightarrow{p} \frac{\mathbb{E}[Y(P - \mathbb{E}[P \mid X])]}{\mathbb{E}[T(P - \mathbb{E}[P \mid X])]} \\ &= \int \sum_{j=1}^{J_x - 1} \frac{\theta_j(x)}{\int \sum_{j=1}^{J_x-1} \theta_j(x) \, \mathrm{d}F^X(x)} \alpha(p_{j,x}; x) \, \mathrm{d}F^X(x),\end{aligned}$$

where (Y, T, P, X) are distributed according to the limiting distribution $F^{Y,T,P,X}$, and

$$\theta_j(x) = (p_{j+1,x} - p_{j,x}) \mathbb{P}(P > p_{j,x} \mid X = x) \left(\mathbb{E}[P \mid X = x, P > p_{j,x}] - \mathbb{E}[P \mid X = x] \right).$$

Thus, UJIVE estimates a convex combination of local average treatment effects. This conclusion is robust to many instruments, many covariates, and heteroscedasticity.

1.7 Conclusion

In this chapter, I derived estimands of estimators based on a classic linear IV model under treatment effect heterogeneity. I assumed that the instruments satisfy the monotonicity condition of Angrist and Imbens (1995), so that for each pair of instrument values, we can identify a local average treatment effect (LATE). If the LATEs for all possible instrument pairs are all equal to each other, then all classic estimators estimate this common local average treatment effect. If the LATEs vary, then, under mild assumptions, estimators in the class of two-step IV estimators estimate the same convex combination of them. This class includes the two-stage least squares estimator (TSLS). The estimand of LIML, however, is different, depends on the reduced-form covariance matrix, and may be outside of the convex hull of the local average treatment effects. This possibility makes LIML unattractive in settings with treatment effect heterogeneity.

Unfortunately, the TSLS estimator is inconsistent under many instrument asymptotics, making it a poor choice of estimator in settings with a large number of instruments. I showed that a different two-step IV estimator, the unbiased jackknife IV estimator (UJIVE), on the other hand, remains consistent for a convex combination of LATEs under a many instrument asymptotic sequence that allows for heteroscedasticity, and lets the number of instruments and covariates increase in proportion with the sample size. I therefore recommend that in settings with many instruments, empirical researchers use UJIVE instead of LIML or TSLS.

Chapter 2

Identification and Inference with Many Invalid Instruments[1]

2.1 Introduction

In this paper we study estimation and inference in settings where the interest is in the effect of a potentially endogenous regressor on some outcome. To allow for the possible endogeneity we exploit the presence of additional variables. These variables have some of the features of conventional instrumental variables, in the sense that they are correlated with the endogenous regressor. However, in contrast to conventional instrumental variables, these variables potentially also have direct effects on the outcome, and thus are "invalid" instruments.

Motivated by the context of our applications we explore the identifying power of a novel assumption that the direct effects of these invalid instruments are uncorrelated with the effects of the instruments on the endogenous regressor. We focus on the case with many instruments, allowing their number to increase in proportion with the sample size as in Kunitomo (1980), Morimune (1983) and Bekker (1994). To accommodate the structure in our applications in which the number of instruments is tied to the number of exogenous

[1] co-written with Raj Chetty, John Friedman, Edward Glaeser, and Guido Imbens

covariates, we also allow the number of exogenous covariates to increase in proportion with the sample size, as in Anatolyev (2011).

We show that the limited-information-maximum-likelihood (liml) estimator is no longer consistent once direct effects are present. On the other hand, the modified-bias-corrected-two-stage-least-squares (mbtsls) estimator remains consistent. This estimator is a modification of the bias-corrected two stage least squares estimator (Nagar, 1959; Donald and Newey, 2001) that allows for many exogenous covariates. The intuition for this result is that the liml estimator attempts to impose proportionality of *all* the reduced form coefficients. On the other hand mbtsls, like the two-stage least squares (tsls) estimator, can be thought of as a two-stage estimator. In the first stage a single instrument is constructed as a function of only instruments and endogenous regressors, not involving the outcome variable. This constructed instrument is then used in the second stage to estimate the parameter of interest using methods for just-identified settings. Identification only requires validity of the constructed instrument, not of all the individual instruments. The robustness of the mbtsls estimator comes at a price: the estimator is less efficient than liml in the absence of these direct effects under normality and homoskedasticity.

We also show that conventional tests for over-identifying restrictions, adapted to the many instruments setting, can be used to test for the presence of these direct effects. We recommend in practice that researchers carry out such tests and compare estimates based on liml and the modified version of bias-corrected tsls. We illustrate in the context of two applications that such practice can be illuminating.

The paper is related to two strands of literature. First, we contribute to the literature on many and weak instruments, started by Kunitomo (1980), Morimune (1983), Bekker (1994), Staiger and Stock (1997), and Chao and Swanson (2005). In recent work Anatolyev (2011) relaxes the assumption of fixed number of exogenous regressors. Hausman *et al.* (2012); Chao, Swanson, Hausman, Newey and Woutersen (2012) and Ackerberg and Devereux (2009) relax the assumption of homoscedasticity. Hansen *et al.* (2008), Belloni *et al.* (2012) and Gautier and Tsybakov (2011) allow the first stage to be estimated non-parametrically. This

paper takes a complementary approach: we relax the assumption of no direct effects, but keep the rest of the model simple to maintain tractability. Our key contribution is to show that the superiority of liml in the homoscedastic normal error case with many instruments is tied to the assumption of no direct effects. The mbtsls estimator is shown to be less efficient than liml in the case with no direct effects, but robust to the presence of uncorrelated direct effects.

Second, we contribute to the literature studying properties of instrumental variables methods allowing for direct effects of the instruments. This literature has largely focused on the case with a fixed number of instruments. The focus of this literature has been on correcting size distortions of tests, biases of estimators, sensitivity analyses, and bounds in the presence of direct effects. Fisher (1961, 1966, 1967), Caner (2007); Berkowitz, Caner and Fang (2008) and Guggenberger (2012) analyze the implications of local (small) violations of exogeneity assumption. Hahn and Hausman (2005) compare biases for different estimators in the presence of direct effects. Conley, Hansen and Rossi (2012); Ashley (2009) and Kraay (2008) propose sensitivity analyses in the presence of possibly invalid instruments. Nevo and Rosen (2012) consider assumptions about the sign of the direct effects of the instruments on the outcome to derive bounds on the parameters of interest. Reinhold and Woutersen (2011) and Flores and Flores-Lagunes (2010) also derive bounds allowing for direct effects of the instruments on the outcome. The current paper is the first to derive (point) identification results in the presence of non-local departures from the no-direct-effects assumption or exclusion restriction.

The rest of the paper is organized as follows. In Section 2.2 we discuss in detail the empirical setting that motivates our study, based on Chetty *et al.* (2011). In Section 2.3 we set up the general problem and formulate the critical assumptions. Next, in Section 2.4 we report on the large sample properties of k-class estimators, which covers both liml and mbtsls. In Section 2.5 we discuss tests for instrument validity. We then analyze two data sets to illustrate the usefulness of the results in Section 2.6. In Section 2.7 we report the results of a small simulation study to assess the accuracy of our asymptotic approximations.

Section 2.8 concludes. Proofs are collected in B.

2.2 Motivating Example

In this section we discuss the empirical application that motivates our set up. The application is based on Chetty *et al.* (2011). Chetty *et al.* (2011) are interested in estimating the effect of early achievement for children, as measured by kindergarten performance, on subsequent outcomes, say first grade scores. Chetty *et al.* (2011) wish to exploit the fact that kindergarten teachers are randomly assigned to classes, generating arguably exogenous variation in kindergarten performance. This suggests using kindergarten teacher or classroom indicators as instruments for kindergarten performance. However, a concern with this strategy is that classes mostly stay together over multiple years during the child's education. As a result, kindergarten classroom/teacher assignment is almost perfectly correlated with first grade classroom/teacher assignment. Therefore, the instrument (kindergarten teacher assignment) may have direct effects on the outcome (first grade performance) through first grade teacher assignment, that is, not mediated through the endogenous regressor (kindergarten performance). However, if first grade teachers are randomly assigned, and thus independent of kindergarten teacher assignment, the direct effect of the instrument on the outcome might reasonably be assumed to be independent of the direct effect of the instrument on the endogenous regressor. We show that this independence assumption has substantial identifying power, and discuss estimation strategies that exploit it. The identifying power of this independence assumption suggests that in applications where there is concern regarding the presence of direct effects of the instruments on the outcome it may be useful to explore whether the substantive argument for their presence also suggests that these effects are independent of the effect of the instrument on the endogenous regressor.

To make this precise, let us discuss a simplified version of the Chetty *et al.* (2011) application in more detail. Let us ignore the presence of any exogenous regressors beyond the intercept. Children are indexed by $i = 1, \ldots, N$. The classroom or cluster variable is $G_i \in \{1, 2, \ldots, N_G\}$, where N_G is the number of clusters or classrooms. The instruments

are the classroom indicators, $Z_{ik} = \mathbf{1}_{G_i=k}$, for $k = 1, \ldots, N_G - 1$, so that the number of instruments is the number of clusters minus one. Following the clustering literature we focus on large sample approximations where the number of units in each cluster is finite and the number of clusters increases proportional to the sample size, $N_G/N \to \alpha_K > 0$, leading to the Bekker-style many-instruments asymptotics. In this simple case the model can be written as

$$Y_i = \delta + \beta X_i + \sum_{k=1}^{N_G-1} \gamma_k Z_{ik} + \epsilon_i, \tag{2.1}$$

$$X_i = \pi_{22} + \sum_{k=1}^{N_G-1} \pi_{12,k} Z_{ik} + \nu_i, \tag{2.2}$$

where Y_i is the outcome (first grade test scores) and X_i is the endogenous regressor (kindergarten performance). The residuals ϵ_i and ν_i are assumed to be independent across individuals, but correlated with each other. The coefficient β on the endogenous regressor is the object of interest. The coefficients on the instruments in the second equation, $\pi_{12,k}$ capture the direct effects on the endogenous regressor. Here they represent the effects of the kindergarten teachers on kindergarten performance. The presence of nonzero coefficients on the instruments in the first equation, denoted by γ_k, is what make the instruments invalid. These coefficients represent the effects of the first grade teachers on the first grade test scores.

Similar to the clustering literature, we view the $\pi_{12,k}$ and γ_k as random variables. In this setting where the instruments are cluster indicators this is equivalent to viewing the cluster effects as random, a common assumption in such settings. An alternative formulation of the model, one which stresses the links to the clustering literature, would be

$$Y_i = \delta + \beta X_i + U_{G_i} + \epsilon_i, \tag{2.3}$$

$$X_i = \pi_{22} + V_{G_i} + \nu_i. \tag{2.4}$$

The random classroom component in the outcome equation in the clustering notation, U_{G_i}, is equal to the coefficient on one of the instruments, γ_{G_i}, and the random classroom component

in the equation for X_i, V_{G_i} is equal to the coefficient on the same instrument in the first stage, π_{12,G_i}. The V_{G_i} represents the effect of kindergarten teachers on the kindergarten performance. The U_{G_i} represents the effect of first grade teachers on the outcome. We focus on the notation and formulation in (2.1)–(2.2) because it stresses links to the literature on many instrumental variables that are helpful in motivating the estimators we consider.

The instruments are not valid in the sense that the standard orthogonality condition for instruments does not hold, holding fixed the $\gamma_1, \ldots, \gamma_{N_G-1}$:

$$\mathbb{E}\left[(Y_i - \delta - X_i\beta)Z_i | Z_i, \gamma_1, \ldots, \gamma_{N_G-1}\right] =$$

$$\mathbb{E}\left[U_{G_i}Z_i |\, Z_i, \gamma_1, \ldots, \gamma_{N_G-1}\right] = \begin{pmatrix} \gamma_1 Z_{i1} \\ \vdots \\ \gamma_{N_G-1} Z_{iN_G-1} \end{pmatrix} \neq 0. \quad (2.5)$$

However, we wish to exploit the random assignment of both kindergarten and first grade teachers. We therefore consider the assumption that the effects of kindergarten teachers on kindergarten performance and the effects of first grade teachers on outcomes are independent:

$$\pi_{12,k} \perp\!\!\!\perp \gamma_k,$$

or, given a normalization of the mean of the γ_k, $E[\pi_{12,k}\gamma_k] = 0$. In terms of the cluster formulation (2.3), the assumption is $U_{G_i} \perp\!\!\!\perp V_{G_i}$. This suggests replacing the orthogonality condition (2.5), which requires each instrument to be valid, with

$$\begin{aligned} \mathbb{E}\left[(Y_i - \delta - X_i\beta)Z_i'\pi_{12}\right] &= \mathbb{E}\left[\mathbb{E}\left[\left(Y_i - \delta - X_i\beta\right)Z_i'\pi_{12}\middle|\, Z_i\right]\right] \quad (2.6) \\ &= \mathbb{E}\left[U_{G_i}V_{G_i}\right] = \mathbb{E}\left[\sum_{k=1}^{N_G-1} \gamma_k \pi_{12,k} Z_{ik}\right] = 0, \end{aligned}$$

which requires the instruments to be valid in an average sense. Here π_{12} is the vector with kth element equal to $\pi_{12,k}$. In a setting with a few instruments this would suggest estimating β as the solution to solving (2.6) with π_{12} replaced by the least squares estimator $\hat{\pi}_{12}$:

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y} - \beta(X_i - \overline{X})\right)Z_i'\hat{\pi}_{12},$$

where $\overline{Y}$ and $\overline{X}$ are sample averages of Y_i and X_i respectively. Solving this for β leads to the standard tsls estimator. However, since the work by Bekker (1994) it is well known that even with valid instruments the tsls estimator is not consistent in settings with many instruments, and thus it is unlikely to be consistent here. This motivates looking for alternative, tsls-like, estimators of the type that have been proposed to deal with many-instrument problems. We do so in the Section 2.4. First, in Section 2.3, we introduce the general set up.

2.3 General Set Up

We consider the following instrumental variables model:

$$\begin{aligned} Y_i &= X_i\beta + W_i'\delta + Z_i'\gamma + \epsilon_i. \\ X_i &= Z_i'\pi_{12} + W_i'\pi_{22} + \nu_i. \end{aligned} \tag{2.7}$$

The first equation relates a scalar outcome Y_i, $i = 1, \ldots, N$, to a potentially endogenous scalar regressor X_i. W_i is a vector of exogenous regressors with dimension L_N (including an intercept), and Z_i is a vector of instruments with dimension K_N. The second equation relates the endogenous regressor X_i to the exogenous regressors W_i and the instruments Z_i. The object of interest is the coefficient β on the endogenous regressor in the outcome equation.

The model (2.7) modifies the conventional many-instruments model (e.g. Bekker, 1994) in two ways. First, and this is the main contribution of the paper, we allow γ to be non-zero, thus allowing for direct effects of the instrument on the outcome. If we restrict $\gamma = 0$, then the exclusion restriction holds, and the instruments are valid. If we leave γ unrestricted, then β, the coefficient of interest, is not identified. In this paper, we will consider assumptions on γ that are weaker than $\gamma = 0$, but that still allow us to identify β, and assess their empirical content. Second, like Anatolyev (2011), we allow the number of exogenous regressors, L_N, to change with the sample size. The motivation for this extension is that often the presence of a large number of instruments is the result of interacting a few basic instruments with many exogenous covariates. For example, in Angrist and Krueger (1991), the basic instruments were three quarter of birth indicators. These were interacted with year of birth and state of

birth indicators to generate a large number of instruments. As the results below show, this second extension does not make a substantial difference for the variance calculations, unless the ratio of the number of exogenous variables to the sample size is large.

Because the number of instruments and the number of exogenous variables change with the sample size, the distribution of some of the random variable also changes with the sample size. To be precise, we should therefore index the random variables and parameters by the sample size N. For ease of notation we drop this index. In the remainder γ and π_{12} will be vectors of dimension K_N, and δ and π_{22} will be vectors of dimension L_N.

Next, we introduce some additional notation. Let $\mathbf{Y}$ be the N-component vector with ith element Y_i, $\mathbf{X}$ the N-component vector with ith element X_i, ϵ the N-component vector with ith element ϵ_i, ν the N-component vector with ith element ν_i, $\mathbf{W}$ the $N \times L_N$ matrix with ith row equal to W_i', and $\mathbf{Z}$ the $N \times K_N$ matrix with ith row equal to Z_i'. Let $\overline{\mathbf{X}} = (\mathbf{X}, \mathbf{W})$ be the full matrix of endogenous and exogenous regressors, let $\overline{\mathbf{Y}} = (\mathbf{Y}, \mathbf{X})$ be the full matrix of endogenous variables, and let $\overline{\mathbf{Z}} = (\mathbf{Z}, \mathbf{W})$ be the full matrix of exogenous variables. Define for an arbitrary $N \times J$ matrix $\mathbf{S}$ the following four $N \times N$ matrices, the projection matrix $\mathbf{P_S}$, the matrix $\mathbf{M_S}$ that projects on the orthogonal complement of $\mathbf{S}$, the diagonal matrix $\mathbf{D_S}$ with diagonal elements equal to those of the projection matrix:

$$\mathbf{P_S} = \mathbf{S}\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}', \qquad \mathbf{M_S} = \mathbf{I} - \mathbf{S}\left(\mathbf{S}'\mathbf{S}\right)^{-1}\mathbf{S}', \qquad \mathbf{D_S} = \text{Diag}(\mathbf{P_S})$$

Following Staiger and Stock (1997), we use the subscript $\perp$ as shorthand for taking residuals after regression on the exogenous regressors $\mathbf{W}$, so $\mathbf{Z}_\perp = \mathbf{M_W}\mathbf{Z}$, $\mathbf{X}_\perp = \mathbf{M_W}\mathbf{X}$, $\mathbf{Y}_\perp = \mathbf{M_W}\mathbf{Y}$, and $\overline{\mathbf{Y}}_\perp = \mathbf{M_W}\overline{\mathbf{Y}}$. We also denote by ι_N the N-dimensional vector of ones.

Define the augmented concentration parameter, the two by two matrix Λ_N:

$$\Lambda_N = \begin{pmatrix} \Lambda_{N.11} & \Lambda_{N,12} \\ \Lambda_{N,12} & \Lambda_{N,22} \end{pmatrix} = \begin{pmatrix} \gamma & \pi_{12} \end{pmatrix}' \mathbf{Z}_\perp'\mathbf{Z}_\perp \begin{pmatrix} \gamma & \pi_{12} \end{pmatrix}. \tag{2.8}$$

The (1,1) element, $\Lambda_{N,11}$, measures the degree of misspecification. In the case with valid instruments, $\gamma = 0$, and thus $\Lambda_{N,11} = \Lambda_{N,12} = 0$ and the only non-zero element of Λ_N is $\Lambda_{N,22}$.

We make the following assumptions. Some of these can be weakened along the lines of Chao and Swanson (2005). We focus on the simplest version of the assumptions and results that allow us to focus on the conceptual contribution of the paper.

Assumption 1 (Instruments and exogenous variables). **(i)** $Z_i \in \mathbb{R}^{K_N}$, $W_i \in \mathbb{R}^{L_N}$, $\epsilon_i \in \mathbb{R}, \nu_i \in \mathbb{R}$, for $i = 1, \ldots, N$, $N = 1, \ldots$ are triangular arrays of random variables with $(Z_i, W_i, \epsilon_i, \nu_i)$, $i = 1, \ldots, N$ exchangeable.

(ii) $\overline{\mathbf{Z}}$ is full column rank with probability one.

This assumption is standard, with a minor adaption to allow for many exogenous variables.

Assumption 2 (Model). **(i)** $(\epsilon_i, \nu_i)' \mid \mathbf{Z}, \mathbf{W}$ are iid with mean zero, positive definite covariance matrix Σ, and finite fourth moments;

(ii) The distribution of $(\epsilon_i, \nu_i)' \mid \mathbf{Z}, \mathbf{W}$ is Normal.

To simplify the derivation of distributional results, we will assume that the structural errors Normally distributed. We do not require Normality for consistency arguments. Recent papers by Chao *et al.* (2012) and Hausman *et al.* (2012) investigate the implications of heteroscedasticity in the setting with many valid instruments, and show that liml loses some of its attractive properties in that case. Our results complement theirs in the sense that our results highlight a different concern with conventional estimators such as liml.

Assumption 3 (Number of instruments and exogenous regressors). For some $0 < \alpha_K < 1$ and $0 \leq \alpha_L < 1$, and $\alpha_K + \alpha_L < 1$

$$K_N/N = \alpha_K + o(N^{-1/2}), \qquad \text{and} \qquad L_N/N = \alpha_L + o(N^{-1/2}).$$

The first part of this assumption is standard in the many-instrument literature, with the exception of the restriction that $\alpha_K > 0$. We rule out $\alpha_K = 0$ to allow for the possibility that the probability limit of $\Lambda_{N,11}$ is positive. This is similar to the clustering literature in which the number of clusters needs to increase with the sample size to achieve point-identification. If the probability limit of $\Lambda_{N,11}$ is equal to zero, we can allow for the possibility that K_N is

fixed and $\alpha_K = 0$. The second part is identical to the corresponding assumption in Anatolyev (2011).

Assumption 4 (Concentration parameter). For some positive semi-definite 2×2 matrix Λ with $\Lambda_{22} > 0$,

$$\Lambda_N / N \xrightarrow{p} \Lambda, \qquad \text{and} \qquad \mathbb{E}\left[\Lambda_N / N\right] \to \Lambda.$$

The first part of Assumption 4 is a natural extension of the assumption underlying the Bekker many-instrument asymptotics. The second part of the assumption strengthens this slightly by also requiring the expectation of the concentration parameter to converge to its probability limit.

Assumption 5 (Zero correlation). $\Lambda_{12} = 0$.

The last assumption is a new and critical assumption. We allow for direct effects of the instruments on the outcome ($\Lambda_{11} > 0$), but assume that these direct effects are uncorrelated with the direct effects of the instruments on the endogenous regressor. This is a strong assumption, and one that needs to be justified on a case-by-case basis. In settings such as the Chetty *et al.* (2011) application we argued (in Section 2.2) that this may be a reasonable assumption.

2.4 The Properties of k-Class Estimators

This section contains the main formal results of the paper. We discuss estimators for β and their large sample properties under the assumptions introduced in the previous section. Some of the results are for general k-class estimators (Nagar, 1959; Theil, 1961, 1971; Davidson and MacKinnon, 1993), and some for four particular estimators in this class. All four have been introduced previously, and are asymptotically equivalent in the conventional setting with a fixed number of valid instruments and a fixed number of

exogenous regressors. Given a scalar k, a k-class estimator for (β, δ) is given by:

$$\begin{pmatrix} \hat{\beta}_k \\ \hat{\delta}_k \end{pmatrix} = \left(\overline{\mathbf{X}}' (\mathbf{I} - k\mathbf{M}_{\overline{\mathbf{Z}}}) \overline{\mathbf{X}} \right)^{-1} \left(\overline{\mathbf{X}}' (\mathbf{I} - k\mathbf{M}_{\overline{\mathbf{Z}}}) \mathbf{Y} \right).$$

We are primarily interested in the estimator for β, which can be written using the $\perp$ projection notation as

$$\hat{\beta}_k = \left(\mathbf{X}_\perp' (\mathbf{I} - k\mathbf{M}_{\mathbf{Z}_\perp}) \mathbf{X}_\perp \right)^{-1} \left(\mathbf{X}_\perp' (\mathbf{I} - k\mathbf{M}_{\mathbf{Z}_\perp}) \mathbf{Y}_\perp \right). \tag{2.9}$$

A prominent member of the k-class is the two-stage-least-squares (tsls) estimator (Basmann, 1957; Theil, 1961), with $\hat{k}_{\text{TSLS}} = 1$. Even if all instruments are valid, this estimator has been shown to be inconsistent under many-instrument asymptotics, see Kunitomo (1980) and Bekker (1994). We also consider a bias-corrected version of the tsls estimator that is valid under many-instrument asymptotics. Nagar (1959) suggested the bias correction $\hat{k}_{\text{nagar}} = 1 + (K_N - 2)/N$, but the second of the four estimators we focus on is a slightly different version suggested by Donald and Newey (2001), with

$$\hat{k}_{\text{BTSLS}} = \frac{1}{1 - (K_N - 2)/N}.$$

Although in samples with a moderate number of instruments the difference between the Nagar and Donald-Newey estimators is small, this difference does not go away under many-instruments asymptotics with $K_N/N \to \alpha_K > 0$, and only the Donald-Newey version is consistent under those asymptotics. Once we also allow L_N to increase with sample size, $\hat{\beta}_{\text{btsls}}$ is no longer consistent. To address this issue, the third estimator we consider is a further modification of the Donald-Newey bias-corrected estimator, first suggested by Anatolyev (2011), that is consistent even when $L_N/N \to \alpha_L > 0$:

$$\hat{k}_{\text{MBTSLS}} = \frac{1 - L_N/N}{1 - K_N/N - L_N/N}.$$

In practice this modification has only a minor effect, unless the ratio of the number of exogenous variables to the sample size is substantial.

The fourth estimator we consider is the limited-information-maximum-likelihood (liml) estimator Anderson and Rubin (1949), with

$$\hat{k}_{\text{LIML}} = \min_{\beta} \frac{(\mathbf{Y} - \mathbf{X}\beta)' \mathbf{M}_{\mathbf{W}} (\mathbf{Y} - \mathbf{X}\beta)}{(\mathbf{Y} - \mathbf{X}\beta)' \mathbf{M}_{\overline{\mathbf{Z}}} (\mathbf{Y} - \mathbf{X}\beta)}.$$

This estimator has been shown to be asymptotically efficient under many-instrument asymptotics (Chioda and Jansson, 2009; Anderson *et al.*, 2010) in the class of invariant estimators given normality and homoskedasticity of the error terms.

The first of our two main results describes the probability limit of a general k-class estimator under the assumptions given in the previous section.

Theorem 2.1 (Probability limits of k-class estimators). *Suppose Assumptions 1, 2(i), 3, 4 and 5 hold. If* $\hat{k} \xrightarrow{p} k$ *with* $k < \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{\Lambda_{22}}{\Sigma_{22}(1-\alpha_K-\alpha_L)}$*, then:*

$$\hat{\beta}_{\hat{k}} \xrightarrow{p} \beta_k = \beta + \frac{(1 - \alpha_L - (1 - \alpha_K - \alpha_L)k)\Sigma_{12}}{\Lambda_{22} + (1 - \alpha_L - (1 - \alpha_K - \alpha_L)k)\Sigma_{22}}.$$

If we impose $\alpha_L = 0$, the condition for consistency of $\hat{\beta}_{\hat{k}}$ is the same as in Chao and Swanson (2005), namely that $\hat{k} \to 1/(1 - \alpha_K)$. Having many exogenous regressors changes the condition on $\hat{k}$ to $\hat{k} \to (1 - \alpha_L)/(1 - \alpha_K - \alpha_L)$. As long as $\Lambda_{12} = 0$, this result holds whether or not $\Lambda_{11} > 0$. Therefore, the robustness of a *k*-class estimators to the presence direct effects depends on whether the probability limit of $\hat{k}$ remains unaffected by their presence.

For the four estimators we discussed, the implication of this theorem is given in the following Corollary.

Corollary 2.1. *Suppose Assumptions 1, 2(i), 3, 4 and 5 hold. Then:*

(i) *(tsls)*

$$\beta_{\text{TSLS}} = \beta + \frac{(1 - \alpha_L - (1 - \alpha_K - \alpha_L))\Sigma_{12}}{\Lambda_{22} + (1 - \alpha_L - (1 - \alpha_K - \alpha_L))\Sigma_{22}}, \quad k_{\text{TSLS}} = 1,$$

(ii) *(btsls)*

$$\beta_{\text{BTSLS}} = \beta + \frac{\{\alpha_K \alpha_L/(1-\alpha_K)\}\,\Sigma_{12}}{\Lambda_{22} + \{\alpha_K \alpha_L/(1-\alpha_K)\}\,\Sigma_{22}}, \quad k_{\text{BTSLS}} = \frac{1}{1-\alpha_K},$$

(iii) *(mbtsls)*

$$\beta_{\text{MBTSLS}} = \beta, \quad k_{\text{MBTSLS}} = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L},$$

(iv) *(liml) Suppose* $\min \text{eig}(\Sigma^{-1}\Lambda) < \Lambda_{22}/\Sigma_{22}$. *Then:*

$$\beta_{\text{LIML}} = \beta - \frac{\min \text{eig}(\Sigma^{-1}\Lambda)\Sigma_{12}}{\Lambda_{22} - \min \text{eig}(\Sigma^{-1}\Lambda)\Sigma_{22}}, \quad k_{\text{LIML}} = \frac{1-\alpha_L}{1-\alpha_K-\alpha_L} + \frac{\min \text{eig}(\Sigma^{-1}\Lambda)}{1-\alpha_K-\alpha_L},$$

The key insight is that the mbtsls modification of the tsls estimator that makes it robust to the presence of many instruments and many exogenous variables is also robust to the presence of direct effects, provided these direct effects are uncorrelated with the effects of the instrument on the endogenous regressor. On the other hand, in order for liml to be consistent for all values of Σ, then it has to be the case that Λ_{11} is equal to zero since $\min \text{eig}(\Sigma^{-1}\Lambda) > 0$ otherwise. To provide some intuition, consider the reduced-form based on the model (2.7):

$$\begin{aligned} Y_i &= Z_i'(\pi_{12}\beta + \gamma) + W_i'(\delta + \pi_{22}\beta) + (\nu_i\beta + \epsilon_i), \\ X_i &= Z_i'\pi_{12} + W_i'\pi_{22} + \nu_i. \end{aligned}$$

If the instruments are valid, so that $\gamma = 0$, then the vector of reduced-form coefficients on Z_i in the first equation is proportional to π_{12}, the vector of reduced-form coefficients in the second equation. The liml estimator tries to impose this proportionality. This leads to efficiency if proportionality holds, under normality and homoskedasticity, (Chioda and Jansson, 2009; Anderson *et al.*, 2010). However, if $\gamma \neq 0$, then the proportionality does not hold in the population, and liml loses consistency. On the other hand, mbtsls and mjive, like tsls, can be thought of as two stage estimators. In the first stage composite instruments are constructed, one for each regressor (endogenous or exogenous) based on

the data on the endogenous regressor, the exogenous variables, and the instruments alone. These instruments are then used to estimate the parameters of interest using a method for just-identified settings, possibly with some adjustment. In this procedure proportionality of the reduced forms is never exploited. This explains why $\Lambda_{12} = 0$ is a sufficient condition for consistency, although it results in efficiency loss relative to liml when proportionality does hold.

Without the assumption that the direct effects are uncorrelated (Assumption 5), the probability limit of k-class estimators has an additional term that is proportional to Λ_{12}:

$$\hat{\beta}_{\hat{k}} \xrightarrow{p} \beta + \frac{\Lambda_{12} + (1 - \alpha_L - (1 - \alpha_K - \alpha_L)k)\Sigma_{12}}{\Lambda_{22} + (1 - \alpha_L - (1 - \alpha_K - \alpha_L)k)\Sigma_{22}} \tag{2.10}$$

In this case all the k-class estimators are in general inconsistent, and in fact there are no estimators for β that are consistent for all values of Σ.

Note also that the bias of the btsls estimator is minor if $\Lambda_{12} = 0$: it is essentially proportional to the product of α_K and α_L, so that unless both are substantial, the bias will generally be small. However, the presence of many exogenous regressors might have a large effect on the probability limits of other estimators. For example, in previous version of this paper (Kolesár *et al.*, 2011) we show that the jackknife instrumental variables estimator (Angrist *et al.*, 1999) may exhibit substantial bias when the number of exogenous covariates is large.

The second main result concerns the asymptotic approximation to the distribution of the mbtsls estimator. We focus on the mbtsls estimator because that is the only estimator in the k-class that is consistent under the assumptions we consider. A complication arises because, except in the special case where the only non-zero element of Λ_N is the $(2,2)$ element $\Lambda_{N,22}$ (the standard case with valid instruments, $\Lambda_{11} = 0$), the asymptotic distribution for $\hat{\beta}_{\text{mbtsls}}$ depends on the stochastic properties of $\Lambda_N - \Lambda$. In order to derive the asymptotic distribution of $\hat{\beta}_{\text{mbtsls}}$ we therefore make one additional assumption about the sequence of γ_k and $\pi_{12,k}$. That is, similar to corresponding assumptions in the clustering literature, we assume that these parameters are random and make assumptions regarding their stochastic

properties. First we redefine the parameters by orthogonalizing them with respect to $\mathbf{Z}_\perp$ as

$$\begin{pmatrix} \tilde{\gamma} & \tilde{\pi}_{12} \end{pmatrix} = \left(\alpha_K \mathbf{Z}_\perp' \mathbf{Z}_\perp\right)^{1/2} \begin{pmatrix} \gamma & \pi_{12} \end{pmatrix}.$$

Assumption 6 (Incidental parameters). The pairs $(\tilde{\gamma}_k, \tilde{\pi}_{12,k})$, for $k = 1, 2, \ldots, K_N$, are iid with distribution

$$\begin{pmatrix} \tilde{\gamma}_k \\ \tilde{\pi}_{12,k} \end{pmatrix} \mid \mathbf{Z}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\gamma \\ \mu_\pi \end{pmatrix}, \Xi \right).$$

The motivation for formulating the random effects assumption in terms of the orthogonalized parameters rather than in terms of the original parameters comes from the cluster structure in our application where the instruments are indicators for the clusters. Exploiting that special structure the augmented concentration parameter can be written as the sample covariance matrix of $(\gamma_k, \pi_{12,k})$:

$$\Lambda_N = \frac{N}{N_G} \sum_{k=1}^{N_G - 1} \begin{pmatrix} (\gamma_k - \overline{\gamma})^2 & (\gamma_k - \overline{\gamma})(\pi_{12,k} - \overline{\pi}_{12}) \\ (\gamma_k - \overline{\gamma})(\pi_{12,k} - \overline{\pi}_{12}) & (\pi_{12,k} - \overline{\pi}_{12})^2 \end{pmatrix},$$

where

$$\overline{\gamma} = \frac{1}{N_G} \sum_{k=1}^{N_G - 1} \gamma_k, \qquad \text{and} \qquad \overline{\pi}_{12} = \frac{1}{N_G} \sum_{k=1}^{N_G - 1} \pi_{12,k}.$$

Now let us consider Assumption 6 and interpret it in this context. Suppose we have a large population of clusters. Let $\delta + U_k$ and $\pi_{22} + V_{X,k}$ be the population means of $Y_i - \beta X_i$ and X_i in cluster k, and let δ and π_{22} be the population average of the cluster means. In terms of the original parametrization, we have: $\pi_{12,k} = V_k$, and $\gamma_k = U_k$.

The natural way to impose a random effects structure on the parameters would be to assume that the cluster means $(\delta + U_k, \pi_{22} + V_k)$ are independent and normally distributed:

$$\begin{pmatrix} \delta + U_k \\ \pi_{22} + V_k \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \delta \\ \pi_{22} \end{pmatrix}, \Phi \right). \tag{2.11}$$

This implies

$$\begin{pmatrix} \tilde{\gamma} & \tilde{\pi}_{12} \end{pmatrix} = \sqrt{\frac{N_G - 1}{N_G}} B \begin{pmatrix} \delta + U_1 & \pi_{22} + V_1 \\ \vdots & \vdots \\ \delta + U_{N_G} & \pi_{22} + V_{N_G} \end{pmatrix},$$

$$B = \left(\begin{array}{c|c} I_{N_G-1} - \frac{1-1/\sqrt{N_G}}{N_G-1} \iota_{N_G-1} \iota'_{N_G-1} & -\frac{1}{\sqrt{N_G}} \iota_{N_G-1} \end{array} \right)$$

where the $(N_G - 1) \times N_G$ matrix B satisfies $B\iota_{N_G} = 0$, and $BB' = \mathbf{I}_{N_G-1}$. Thus, a random effects specification on $(\delta + U_k, \pi_{22} + V_k)$ as in (2.11) implies a random effects specification on $(\tilde{\gamma}, \tilde{\pi}_{12})$, namely

$$\begin{pmatrix} \tilde{\gamma}_k \\ \tilde{\pi}_{12,k} \end{pmatrix} \Bigg| \mathbf{Z}, \mathbf{W} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Xi \right), \qquad \text{with } \Xi = \frac{N_G - 1}{N_G} \cdot \Phi.$$

Given Assumption 6 it follows that the augmented concentration parameter satisfies

$$\begin{aligned} \Lambda &= \operatorname{plim}\left(\frac{\Lambda_N}{N}\right) = \operatorname{plim}\left(\frac{1}{N} \begin{pmatrix} \gamma' \\ \pi'_{12} \end{pmatrix} \mathbf{Z}'_\perp \mathbf{Z}_\perp \begin{pmatrix} \gamma & \pi_{12} \end{pmatrix} \right) \\ &= \operatorname{plim}\left(\frac{1}{K_N} \sum_{k=1}^{K_N} \begin{pmatrix} \tilde{\gamma}_k \\ \tilde{\pi}_{12,k} \end{pmatrix} \begin{pmatrix} \tilde{\gamma}_k & \tilde{\pi}_{12,k} \end{pmatrix} \right) = \begin{pmatrix} \mu_\gamma \\ \mu_\pi \end{pmatrix} \begin{pmatrix} \mu_\gamma \\ \mu_\pi \end{pmatrix}' + \Xi. \end{aligned}$$

Now we can state the second main result of the paper.

Theorem 2.2 (Asymptotic normality with many invalid instruments). *Suppose that Assumptions 1–6 hold. Suppose in addition that $\mu_\gamma = \Xi_{12} = 0$. Then*

$$\begin{aligned} &\sqrt{N}\left(\hat{\beta}_{\text{MBTSLS}} - \beta\right) \Rightarrow \\ &\qquad \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right) + \Lambda_{11}\left(\Sigma_{22} + \frac{\Lambda_{22}}{\alpha_K}\right)\right)\right). \end{aligned} \tag{2.12}$$

Note that here we do assume normality (Assumption 2(ii)).

If in addition $\Lambda_{11} = 0$ (corresponding to the conventional many-instrument case), the

distribution for $\hat{\beta}_{\text{MBTSLS}}$ is the special case of (2.12) with $\Lambda_{11} = 0$:

$$\sqrt{N}\left(\hat{\beta}_{\text{MBTSLS}} - \beta\right) \mid \overline{\mathbf{Z}} \Rightarrow \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2\right)\right)\right). \tag{2.13}$$

In this case imposing Assumption 6 has no effect on the asymptotic distribution. This result obtains because under the standard valid many-instrument asymptotic sequence, the Normal prior on the incidental parameters gets dominated, and the Bernstein-von Mises theorem applies (see Kolesár, 2012).

The asymptotic variance of $\hat{\beta}_{\text{MBTSLS}}$ is strictly larger if $\Lambda_{11} > 0$ than if $\Lambda_{11} = 0$. The additional term in the variance, $\Lambda_{11}\left(\Sigma_{22} + \Lambda_{22}/\alpha_K\right)$, diverges if α_K goes to zero. As often in settings with clustering, the number of clusters needs to increase proportional to the sample size for convergence of the estimator to be at $\sqrt{N}$ rate. In contrast to much of the many-instruments literature, the presence of many instruments is required here for consistency, rather than being a nuisance.

For comparison, the asymptotic distribution of liml given $\Lambda_{11} = 0$ is

$$\sqrt{N}\left(\hat{\beta}_{\text{LIML}} - \beta\right) \mid \overline{\mathbf{Z}} \Rightarrow \mathcal{N}\left(0, \Lambda_{22}^{-2}\left(\Sigma_{11}\Lambda_{22} + \frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}\left(\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2\right)\right)\right), \tag{2.14}$$

with a smaller variance than the mbtsls estimator under the same assumptions (comparing (2.13) with (2.14)), consistent with the efficiency of liml under those conditions. There is therefore a trade-off between the robustness of the mbtsls estimator to the presence of direct (uncorrelated) effects and the efficiency of liml in the absence of such effects (under normality and homoskedasticity).

2.5 Testing

The assumption that the instruments are valid (that is, that $\gamma = 0$) is equivalent to restricting the $\Lambda_{11,N}$ (and thus $\Lambda_{12,N}$) elements of the augmented concentration matrix to zero. Several tests of this restriction have been proposed in the literature, most of them in the setting with a fixed number of instruments, but some designed to be robust to the presence of many instruments.

The most popular one test, due to Sargan (1958), is based on the statistic:

$$J_{\text{Sargan}} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{LIML}})'\mathbf{P}_{\mathbf{Z}_\perp}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{LIML}})}{(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{LIML}})'\mathbf{M}_{\mathbf{W}}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\text{LIML}})/N} = N(1 - \hat{k}_{\text{LIML}}^{-1})$$

This statistic can easily be computed as the $N \cdot R^2$ from regressing the estimated residuals in the structural equation on instruments and exogenous regressors. Sargan (1958) shows that under the standard strong instrument asymptotic sequence which keeps the number of instruments and exogenous regressors fixed (so that $K_N = K$ and $L_N = L$), this statistic satisfies $J_{\text{Sargan}} \Rightarrow \chi^2_{K-1}$. Anatolyev and Gospodinov (2011) show, however, that if the number of instruments is allowed to grow with the sample size, the limiting distribution is Normal, and using a critical value based on the χ^2 distribution with $K_N - 1$ degrees of freedom yields an asymptotically conservative test. Anatolyev and Gospodinov (2011) therefore propose an adjustment to the critical value. Unfortunately, if the number of exogenous regressors is allowed to grow with the sample size as well, the original as well as the adjusted Sargan test have asymptotic size equal to one Anatolyev (2011). We therefore propose to use a test statistic suggested by Cragg and Donald (1993):

$$J_{\text{Cragg-Donald}} = (N - K_N - L_N)(\hat{k}_{\text{LIML}} - 1)$$

Like the Sargan statistic, this statistic depends on the data only through $\hat{k}_{\text{LIML}}$. Both tests reject for large values of $\hat{k}_{\text{LIML}}$, so their power properties are identical; the only difference between them is in how well they control size. Under the standard strong instrument asymptotics, this statistic, like the Sargan statistic, is also distributed according to χ^2_{K-1}. However, under many-instrument asymptotics, using the $1 - \tilde{\alpha}$ quantile of the χ^2 distribution with $(K_N - 1)$ degrees of freedom for a test with nominal size $\tilde{\alpha}$ results in asymptotic size distortions. We therefore compare $J_{\text{Cragg-Donald}}$ against the $\Phi(\sqrt{(1-\alpha_L)/(1-\alpha_K-\alpha_L)}\Phi^{-1}(1-\tilde{\alpha}))$ quantile of $\chi^2_{K_N-1}$, where Φ is the cdf of a standard Normal distribution. Kolesár (2012) shows that this adjusted Cragg-Donald test controls size under strong, as well as many-instrument asymptotics.

2.6 Two Applications

In this section we discuss two applications. These will serve to provide further context for the empirical content of the assumptions, and in particular the zero correlation assumption (Assumption 5).

2.6.1 Application I

The first application is based on Chetty *et al.* (2011) first introduced in Section 2.2. The interest in Chetty *et al.* (2011) is in the effect of kindergarten performance on later outcomes. Here we focus on first, second, and third grade performance as the outcome of interest. The outcome equation is

$$Y_i = \beta X_i + \sum_{\ell=1}^{L_N} \delta_\ell W_{i\ell} + \sum_{k=1}^{K_N} \gamma_k Z_{ik} + \epsilon_i. \tag{2.15}$$

Here the outcome Y_i is first, second, or third grade performance. The endogenous regressor X_i is kindergarten performance. The exogenous regressors W_{ik} include 76 school indicators and three demographic variables (female, black, and being on subsidized lunches), for a total of $L_N = 79$ exogenous variables. The instruments are $K_N = 238$ classroom indicators. The first stage is

$$X_i = \sum_{k=1}^{K_N} \pi_{12,k} Z_{ik} + \sum_{\ell=1}^{L_N} \pi_{22,\ell} W_{i\ell} + \nu_i. \tag{2.16}$$

The motivation for the zero correlation assumption is that the γ_k represent the effects of the first or subsequent, grade teachers. Because the classes largely stay the same from year to year, children with the same kindergarten teacher would have the same first, second, and third grade teacher. However, by design the subsequent teachers were assigned randomly, independently of the kindergarten teachers, and so the γ_k would be independent of the $\pi_{12,k}$ if the only direct effect of the kindergarten classroom/teacher assignment was through the subsequent teacher.

Finally, we impose a random effects structure on the effects of the instruments on

outcomes and endogenous regressors:

$$\left(\begin{array}{c} \tilde{\gamma}_k \\ \tilde{\pi}_{12,k} \end{array} \right) \Bigg| \mathbf{Z}, \mathbf{W} \sim \mathcal{N} \left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \Lambda \right).$$

where, as before, the $(\tilde{\gamma}_k, \tilde{\pi}_{12,k})$ are the orthogonalized coefficients on the instruments:

$$\left(\tilde{\gamma} \quad \tilde{\pi}_{12} \right) = \left(\alpha_K \mathbf{Z}_\perp' \mathbf{Z}_\perp \right)^{1/2} \left(\gamma \quad \pi_{12} \right).$$

In Table 2.1 we present point estimates of the parameter of interest, β, based on tsls, liml, btsls, and mbtsls. For each of the estimators we present up to four different standard errors: conventional standard errors, Bekker standard errors which are robust to the presence of many instruments, standard errors robust to the presence of many instruments and many exogenous regressors, and standard errors robust to the presence of direct effects of the instruments on the outcome. For all three outcomes the liml estimate differ substantially from tsls. Based on the early many-instrument literature one might interpret that as evidence of the bias of the tsls estimator in settings with many instruments, and view the liml estimates are more credible. However, the btsls and mbtsls estimates, which, like liml, would be consistent under the conventional many-instruments asymptotics, also differ substantially from the liml estimates.

To understand the difference between the liml and btsls/mbtsls estimates, we report in Table 2.2 test statistics and p-values for the tests for instrument validity $\Lambda_{11} = 0$. The results from these tests are consistent with substantial variation in the γ_k. Although these results do not validate the mbtsls estimates (for that one still relies on the zero correlation assumption, $\Lambda_{12} = 0$), at the very least they imply that the liml estimates should not be taken at face value.

2.6.2 Application II

In the second application we apply some of the methods to a subset of the Angrist and Krueger (1991) data. We use individuals born in the first and fourth quarter (so we have a

Table 2.1: *Estimates for Chetty et al. (2011) Data ($N = 4,170$)*

		Standard Error			
Estimator	$\hat{\beta}$	classic	bekker	many exo	$\Lambda_{11} > 0$
		Panel I: Grade 1 Test scores			
tsls	0.379	(0.037)			
liml	0.014	(0.046)	(0.051)	(0.051)	
btsls	0.221	(0.041)	(0.051)		
mbtsls	0.214	(0.041)	(0.051)	(0.051)	(0.066)
		Panel II: Grade 2 Test scores			
tsls	0.388	(0.043)			
liml	0.108	(0.049)	(0.056)	(0.056)	
btsls	0.233	(0.046)	(0.058)		
mbtsls	0.225	(0.046)	(0.059)	(0.059)	(0.069)
		Panel III: Grade 3 Test scores			
tsls	0.384	(0.048)			
liml	0.174	(0.051)	(0.061)	(0.061)	
btsls	0.238	(0.050)	(0.063)		
mbtsls	0.230	(0.050)	(0.063)	(0.063)	(0.070)

Table 2.2: *Tests of Null Hypothesis* $\Lambda_{11} = 0$ *for Chetty et al. (2011) Data.*

	Sargan		Craig-Donald	
	Test Statistic	p-value	Test Statistic	p-value
Grade 1 Test scores	382.9	< 0.001	31.5	< 0.001
Grade 2 Test scores	319.3	< 0.001	13.6	< 0.001
Grade 3 Test scores	284.6	0.001	4.45	0.035

single binary basic instrument, although this is not essential), dropping observations from Alaska because there are some years birth quarters with no observations, leaving us with observations on 162,487 individuals.

Let W_{ik}, for $k = 1, \ldots, K_N$ be the cluster indicators, corresponding to year-of-birth times state-of-birth interactions, so that $K_N = 500$, and let Q_i be the binary quarter-of-birth indicator. The general model we consider is

$$Y_i = \beta X_i + \sum_{k=1}^{K_N} \delta_k W_{ik} + \sum_{k=1}^{K_N} \gamma_k Q_i W_{ik} + \epsilon_i, \tag{2.17}$$

$$X_i = \sum_{k=1}^{K_N} \pi_{12,k} Q_i W_{ik} + \sum_{k=1}^{K_N} \pi_{22,k} W_{ik} + \nu_i, \tag{2.18}$$

with the random effects structure

$$\left(\begin{array}{c} \gamma_k \\ \pi_{12,k} \end{array} \right) \Bigg| \, \mathbf{Z}, \mathbf{W} \sim \mathcal{N} \left(\left(\begin{array}{c} \mu_\gamma \\ \mu_\pi \end{array} \right), \Xi \right).$$

The critical assumption that $\Lambda_{12} = 0$ is more difficult to justify in this case than in the Chetty *et al.* (2011) case. Its plausibility relies on the interpretation of the direct effects of the instruments on the outcome and the endogenous regressor. The argument for the direct effects of the instrument on the endogenous regressor in the AK study is that quarter of birth effects years of schooling through compulsory schooling laws. If the direct effects of

Table 2.3: *Estimates for Angrist and Krueger (1991) Data ($N = 162,487$)*

		Standard Error			
Estimator	$\hat{\beta}$	classic	bekker	many exo	$\Lambda_{11} > 0$
tsls	0.073	(0.017)			
liml	0.095	(0.017)	(0.042)	(0.042)	
btsls	0.097	(0.017)	(0.039)		
mbtsls	0.098	(0.017)	(0.040)	(0.040)	(0.039)

quarter of birth on earnings is through other differences between states, either in institutions or in economic climate, it may be reasonable to assume that these other differences are uncorrelated with compulsory schooling laws. However, unlike in the Chetty *et al.* (2011) study, there is no design feature that makes this assumption more plausible. Nevertheless, in our view it is still useful to calculate both liml and mbtsls, and calculating the p-value for the test of instrument validity. Finding that the estimators are similar, and that the p-values are not unusually small, lends support to the instrumental variables estimates.

We report in Table 2.3 estimates for β based on tsls, liml, btsls, and mbtsls and the various standard errors. In Table 2.4 we report the results based on the Sargan and Craig-Donald tests for validity of instruments. Here we find, in contrast to the findings for the Chetty *et al.* (2011) data, that the three estimators, liml, btsls, and mbtsls are very similar, and that there is no evidence of direct effects of the instruments on the outcome. Note also that although K_N and L_N are equal in magnitude, the additional adjustment in moving from btsls to mbtsls again makes little difference.

Table 2.4: *Tests of Null Hypothesis* $\Lambda_{11} = 0$ *for Angrist and Krueger (1991) Data*

	Sargan		Craig-Donald	
	Test Statistic	p-value	Test Statistic	p-value
log earnings	487.0	0.64	0.21	0.64

2.7 A Simulation Study

We also carried out a small simulation study to assess the finite sample properties of the estimators. The design was based on the Chetty *et al.* (2011) study. The model is

$$Y_i = \beta X_i + \sum_{\ell=1}^{L_N} \delta_\ell W_{i\ell} + \sum_{k=1}^{K_N} \gamma_k Z_{ik} + \epsilon_i, \tag{2.19}$$

$$X_i = \sum_{k=1}^{K_N} \pi_{12,k} Z_{ik} + \sum_{\ell=1}^{L_N} \pi_{22,\ell} W_{i\ell} + \nu_i, \tag{2.20}$$

where $W_{i\ell}$ and Z_{ik} are school and classroom indicators from the Chetty *et al.* (2011) data, so that $L_N = 76$ and $K_N = 238$. We put a random effects structure on the orthogonalized parameters:

$$\left(\begin{array}{c} \tilde{\gamma}_k \\ \tilde{\pi}_{12,k} \end{array} \right) \Bigg| \, \mathbf{Z}, \mathbf{W} \sim \mathcal{N} \left(\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \Lambda \right).$$

The sample size in the simulations is $N = 4,170$, corresponding to the sample size in the Chetty *et al.* (2011) data, so that $\alpha_L = 0.0182$ and $\alpha_K = 0.0571$.

The values of the parameters are $\delta_j = 0$ and $\pi_{22,j} = 0$, for $j = 1, \ldots, L_N$. The covariance matrix for the structural errors is

$$\Sigma = \left(\begin{array}{cc} 1.0 & 0.5 \\ 0.5 & 1.0 \end{array} \right).$$

The γ_k and π_k are drawn from Normal distributions centered at zero and variances so that

$$\Lambda_{11,N}/K_N = 0.7, \qquad \Lambda_{22,N}/K_N = 2.4,$$

comparable to the values from Chetty *et al.* (2011). We also consider $\Lambda_{11,N} = 0$.

For each of the four estimators we calculate the bias as the average difference between the estimate and the true value (note that liml does not have finite moments, so the bias is arguably not a useful summary measure), the median absolute deviation, and coverage rates based on confidence intervals using the four different standard errors: conventional standard errors, Bekker standard errors which are robust to the presence of many instruments, standard errors robust to the presence of many instruments and many exogenous regressors, and standard errors robust to the presence of direct effects of the instruments on the outcome.

The simulation results are reported in Table 2.5 for the case with valid instruments ($\Lambda_{11} = 0$). In the case with valid instruments, liml performs best, consistent with its efficiency properties. The mbtsls and btsls estimators do almost, but not quite as well. The bekker standard errors do well, the adjustment for many exogenous variables makes virtually no difference for coverage. The tsls estimator performs poorly, not surprising given the presence of many instruments.

When we simulate data with $\Lambda_{11} > 0$ and the instruments are not valid, the results change considerably. The liml estimator now performs very poorly. It has substantial bias and the coverage rates are low. Both the btsls and mbtsls estimators do well in terms of bias and median absolute deviation. Adjusting the variance for the presence of many exogenous covariates makes little difference, but the adjustment to allow for the presence of direct effects makes a considerable difference.

Table 2.5: *Simulations: Coverage Rates for Nominal 95% Confidence Intervals for Different Estimators and Different Standard Errors. 20,000 draws.*

			Standard Errors			
Estimator	average bias	median absolute deviation	classic	bekker	many exo	$\Lambda_{11} > 0$
		Panel I: $\Lambda_{11} = 0$				
tsls	0.147	0.147	2.5			
liml	0.000	0.032	91.3	95.1	95.1	
btsls	0.005	0.034	88.6	94.9		
mbtsls	-0.001	0.034	88.3	95.0	95.1	95.5
		Panel II: $\Lambda_{11} = 0.7$				
tsls	0.147	0.147	7.9			
liml	-0.182	0.182	11.9	15.1	15.1	
btsls	0.005	0.047	76.4	86.1		
mbtsls	-0.001	0.048	76.1	86.1	86.1	93.9

2.8 Conclusion

In this paper we analyze settings with many instruments where each separate instrument might have a direct effect on the outcome. We show that liml is particularly sensitive to such direct effects. In contrast, a modified version of the bias-corrected tsls estimator is robust to such direct effects if these direct effects are uncorrelated with the direct effects of the instrument on the endogenous regressor. We argue in the context of some applications that this orthogonality condition has empirical content. In this setting the choice between liml and the mbtsls estimator depends on a trade-off between efficiency and robustness. In practice we recommend that researchers test for the presence of direct effects under the assumption of orthogonality of the direct effects, and that they compare liml and mbtsls estimates.[2,3]

[2]researchers test for the presence of direct effects under the assumption of orthogonality of the direct effects, and that they compare liml and mbtsls estimates

[3]researchers test for the presence of direct effects under the assumption of orthogonality of the direct effects, and that they compare liml and mbtsls estimates

Chapter 3

Random-Effects Approach to Inference with Many Instruments

3.1 Introduction

This paper provides a principled and unified way of doing inference in a linear instrumental variables model with homoscedastic errors in which the number of instruments is potentially large. The presence of a large number of instruments creates an incidental parameter problem (Neyman and Scott, 1948) because the number of first-stage coefficients corresponds to the number of instruments. To capture this problem in asymptotic approximations, I follow Kunitomo (1980), Morimune (1983), and Bekker (1994) and employ many instrument asymptotics that allow the number of instruments to increase in proportion with the sample size, thus allowing the number of incidental parameters in the model to diverge to infinity. I focus on the case where each instrument is weak in the Staiger and Stock (1997) sense, but collectively the instruments have substantial predictive power, so that the concentration parameter grows at the same rate as the sample size. I allow the rate of growth of the instruments to be zero, in which case the asymptotics reduce to standard strong instrument asymptotics.

One possible way of dealing with the incidental parameter problem is to simply ignore

it, and base inference on full likelihood of the model. This turns out to work for estimation, but not for testing or construction of confidence sets. In particular, the maximum likelihood estimator of the coefficient on the endogenous regressor, β, known as the limited information maximum likelihood (LIML, Anderson and Rubin, 1949) estimator, remains consistent (Bekker, 1994) under many instrument asymptotics. Moreover, LIML is also efficient among estimators that are invariant to rotations of the instruments if the errors are Normal (Chioda and Jansson, 2009). However, the curvature of the likelihood is too large, and likelihood-based tests and confidence sets suffer from size-distortions.

In this paper, I address the incidental parameter problem directly. My basic result is to show that if the errors are Normally distributed, an invariance property of the model and a Bernstein-von Mises type argument can be used to construct an integrated likelihood, which by design delivers inference procedures that are valid under many-instrument asymptotics, and asymptotically optimal under rotation invariance. I show that this likelihood coincides with the random-effects (RE) likelihood of Chamberlain and Imbens (2004), and that the maximum likelihood estimator of β coincides with LIML. Therefore, a simple and principled way of doing inference is to use LIML with standard errors based on the inverse Hessian of the RE likelihood, which I show has a simple closed form.

I derive this basic result in three steps. The first step is to orthogonalize the first stage coefficients so that the information matrix is block-diagonal in the new parametrization. This helps to separate the problem of inference about the parameter of interest β from that of inference about the nuisance parameters.

The second step is to appeal to the invariance principle to reduce the dimensionality of the model. I decompose the orthogonalized first-stage coefficients into a high-dimensional parameter ω_n on the unit sphere which governs the direction of the coefficients, and a scalar parameter λ_n, proportional to the concentration parameter of Rothenberg (1984), that governs their norm. Under rotation invariance, the parameter ω_n drops out, so that the maximal invariant on the parameter space has fixed dimension even as the number of instruments increases to infinity. Imposing invariance is equivalent to assuming a uniform

prior for ω_n, and the likelihood for the maximal invariant (invariant likelihood) is equivalent to an integrated likelihood which integrates ω_n out using this uniform prior.

Since the invariant model is locally asymptotically Normal (Chioda and Jansson, 2009), inference based on the invariant likelihood will be asymptotically efficient in the class of invariant procedures. Moreira (2009) shows that the maximum invariant likelihood estimator of β in the case when the reduced-form covariance matrix Ω is known coincides with LIMLK. I generalize this result along two dimensions. First, if Ω is not known, then the maximum invariant likelihood estimator coincides with LIML. This equivalence explains why LIML is a consistent and efficient invariant estimator despite being based on the concentrated likelihood which in general does not produce consistent estimators in incidental parameter problems. Second, constraining λ_n to equal to a particular value does not affect the maximum invariant likelihood estimate of β.

This result motivates the third step, to put prior a over λ_n in addition to a prior over ω_n and integrate the likelihood over both priors. This additional prior will not affect the maximum integrated likelihood estimator of β, which will still be LIML. If the prior is suitably chosen, the resulting integrated likelihood will yield simpler inference procedures than those based on the invariant likelihood which involve numerical optimization. Moreover, so long as the prior is not dogmatic, it will get dominated in large samples, so that imposing it will not affect asymptotic validity of inference about β either.

The prior I use is a scaled chi-square prior with an unknown scale parameter. This prior, together with a uniform prior on ω_n is equivalent to the random effects prior on the orthogonalized first-stage coefficients proposed by Chamberlain and Imbens (2004): a Normal prior with zero mean and unknown variance (which corresponds to the scale parameter). Therefore, my approach yields an integrated likelihood that is identical to the RE likelihood. Consequently, the random-effects quasi-maximum likelihood estimator of β coincides with LIML.

This analysis yields new insights into the sources of identification in the instrumental variables model, and I use these insights to relax the basic setup along two dimensions. First,

I use it to derive an estimator that is more efficient than LIML when the assumption that the errors are Normally distributed is dropped. In particular, I use the fact that identification of β in the invariant model comes from restrictions on the first moment of the maximal invariant to build a minimum distance objective function. I show that the RE estimator of the model parameters minimizes this minimum distance objective function with respect to a particular weight matrix. This weight matrix is optimal if the errors in the instrumental variables model are Normally distributed, but not otherwise; using weights proportional the inverse of the asymptotic covariance matrix of the moment conditions yields a more efficient estimator. The validity of standard errors for LIML based on the Hessian of the RE likelihood also depends on the assumption of Normality; standard errors based on the conventional GMM/minimum distance formula are robust to non-normality.

The second extension is to relax the exclusion restriction that instruments have no direct effects on the outcome. I assume instead that the direct effects of the instruments are orthogonal to their effect on the endogenous variable. To motivate this assumption, suppose that we are interested in estimating the effect of early achievement for children, as measured by kindergarten performance, on subsequent outcomes, say first grade scores, as in Chetty *et al.* (2011). We want to exploit the fact that in the Project STAR, teachers were randomly assigned to kindergarten classes, and so we use classroom indicators as instruments for kindergarten performance. Suppose that kindergarten teachers only affect first-grade scores through their effect on kindergarten scores, so that the instrument is valid in this sense. However, since classes mostly stay together in subsequent years, kindergarten teacher assignment will be perfectly correlated with first grade teacher assignment. Therefore, the instrument (kindergarten classroom assignment) may have direct effects on the outcome (first grade performance) through first grade classroom assignment, that is not mediated through the endogenous regressor (kindergarten performance). Yet if first grade teachers are also randomly assigned, and thus independent of kindergarten teacher assignment, the direct effect of the instrument on the outcome might reasonably be assumed to be independent of the direct effect of the instrument on the endogenous regressor.

Kolesár *et al.* (2011) show that the β can still be identified under this weaker assumption. They also show that once such direct effects are present, LIML loses consistency, but a slightly modified version of the bias corrected two stage least squares (Nagar, 1959; Donald and Newey, 2001) estimator, the modified bias corrected two stage least squares (MBTSLS) estimator remains consistent. I use the RE framework to gain insight into these results, and to deliver a principled basis for inference robust to the presence of direct effects.

In particular, in addition to modelling the first-stage coefficients as random, I also model the direct effects as Normally distributed and uncorrelated with the first-stage effects. This uncorrelated random effects (URE) model reduces to the random effects model if the variance of the direct effects is restricted to be zero. If the variance parameter is left completely unrestricted, the URE maximum likelihood estimator of the causal effect coincides with MBTSLS, which explains the robustness of MBTSLS to the presence of direct effects, and provides a maximum likelihood motivation for this estimator. If the variance parameter is restricted to be non-negative, the URE maximum likelihood estimator, which I term the URE estimator, is a mixture between MBTSLS and LIML. If the maximum likelihood estimate for the variance of the direct effects is positive, then the URE estimate coincides with MBTSLS. Otherwise, the likelihood is maximized at the boundary, the maximum likelihood estimate for the variance of the direct effects is zero, and URE estimate coincides with LIML. If direct effects are present, then the non-negativity constraint on the variance parameter will not bind in large samples, and the URE estimator has the same asymptotic distribution as MBTSLS. However, if the exclusion restriction holds and no direct effects are present, the URE estimator achieves a lower asymptotic mean squared error than MBTSLS.

These results make URE an attractive robust choice of estimator. One factor complicating inference is that the asymptotic distribution of URE is non-standard when no direct effects are present since in that case the variance of the direct effects is at the boundary of the parameter space. I adapt a procedure from Andrews (1999) that delivers standard errors with correct asymptotic coverage uniformly over the parameter space. An alternative is to use standard errors based on the inverse Hessian, which yield conservative confidence

intervals when no direct effects are present.

The URE model is also helpful in deriving a specification test that is robust to many instruments. When the number of exogenous regressors is allowed to increase with the sample size, the size of the standard Sargan (1958) specification test converges to one as the sample size grows. In the URE model, a specification test of the exclusion restriction is equivalent to a test of the hypothesis that the variance of the direct effects is zero. This equivalence suggests using a test first proposed by Cragg and Donald (1993), but with an adjusted critical value. The adjustment ensures that the test is valid under strong as well as many instrument asymptotics.

This paper draws on two separate strands of literature. First is the literature on many instruments that builds on the on the work by Kunitomo (1980), Morimune (1983), Bekker (1994) and Chao and Swanson (2005). Like Anatolyev (2011), I relax the assumption that the number of exogenous regressors is fixed, and I allow them to grow with the sample size. Hahn (2002), Chamberlain (2007), Chioda and Jansson (2009), and Moreira (2009) focus on optimal inference with many instruments when the errors are Normal and homoscedastic, and my optimality results build on theirs. An interesting new development is to employ shrinkage techniques to obtain more efficient estimators (see, for example, Belloni *et al.*, 2012, Gautier and Tsybakov, 2011, or Carrasco, 2012), although these results rely on an additional sparsity assumption on the first-stage coefficients. Papers by Hansen *et al.* (2008), Anderson *et al.* (2010) and van Hasselt (2010) relax the Normality assumption. Hausman *et al.* (2012), Chao *et al.* (2012), Chao, Hausman, Newey, Swanson and Woutersen (2010) and Bekker and Crudu (2012) also allow for heteroscedasticity. The results for estimation in the presence of direct effects extend those in Kolesár *et al.* (2011).

The second strand of literature is the literature on incidental parameters started by the seminal paper of Neyman and Scott (1948). Lancaster (2000) and Arellano (2003) discuss the incidental parameter problem in a panel data context. Chamberlain and Moreira (2009) relate invariance and random effects approaches to the incidental parameters problem in a dynamic panel data model. My results on the relationship between these two approaches in

an instrumental variables model build on theirs. Sims (2000) proposes a similar random-effects solution in a dynamic panel data model. Moreira (2009) proposes to use the invariance principle. Lancaster (2002) proposes to put a flat prior on the orthogonalized nuisance parameters, rather than the Normal prior with finite unknown variance used here. Cox and Reid (1987) suggest conditioning the likelihood on a maximum likelihood estimate of the orthogonalized incidental parameters. In the instrumental variables model, both proposals yield the concentrated limited information likelihood, and therefore don't deliver valid inference.

The remainder of this paper is organized as follows. Section 3.2 sets up the instrumental variables model, introduces the notation, and finds an orthogonal reparametrization. Section 3.3 uses invariance and Bernstein-von Mises arguments to derive the random effects likelihood and study its properties. Section 3.4 relaxes the Normality assumption and considers a minimum distance approach to inference. Section 3.5 relaxes the exclusion restriction and studies the uncorrelated random effects model. Section 3.6 studies tests of overidentifying restrictions. Section 3.7 concludes. Proofs and derivations are collected in Appendix C.

Notation I denote a d-variate Gaussian distribution with mean μ and variance V by $\mathcal{N}_d(\mu, V)$. $\mathcal{W}_d(\nu, V, M)$ denotes a d-dimensional Wishart distribution with scale matrix V, non-centrality parameter M and ν degrees of freedom, so that if $X_i \sim \mathcal{N}_d(\mu_i, V)$, then $\sum_{i=1}^{\nu} X_i X_i' \sim \mathcal{W}_d(\nu, V, \sum_{i=1}^{\nu} \mu_i \mu_i')$. If $M = 0$, so that the distribution is central Wishart, I omit the last argument and write $\mathcal{W}_d(\nu, V)$. I denote the vector $(1, 0)'$ by e_1, and the vector $(0, 1)'$ by e_2.

3.2 Setup

In this section, I first introduce the model, notation, and the many instrument asymptotic sequence that allows both the number of instruments and the number of exogenous regressors to increase in proportion with the sample size. Second, I reduce the data to the sufficient statistics and find an orthogonal reparametrization of the first-stage coefficients.

Third, I review the failure of the model likelihood, called the limited information likelihood, to deliver asymptotically valid inference.

3.2.1 Model and Assumptions

I consider a linear instrumental variables model with a single endogenous regressor. The model consists of two equations:

$$y_i = x_i\beta + w_i'\delta_n + z_i'\gamma_n + \epsilon_i, \tag{3.1a}$$

$$x_i = z_i'\pi_{12,n} + w_i'\tilde{\pi}_{22,n} + v_{2i}. \tag{3.1b}$$

Equation (3.1a) is a structural equation. The parameter of interest is β, which governs the causal effect of the potentially endogenous regressor x_i on the outcome y_i, $i = 1, \ldots, n$. All remaining parameters in the model are nuisance parameters. w_i is an ℓ_n-dimensional vector of exogenous regressors. I refer to Equation (3.1b) as the first-stage equation. It relates x_i to the exogenous regressors and a k_n-dimensional vector of instruments z_i. The identifying assumption in the model is that the instruments do not appear directly in the structural equation:

Assumption ER (Exclusion restriction). $\gamma_n = 0$.

If $k_n > 1$, then the model is overidentified in the sense that Assumption ER is testable. I discuss tests of the exclusion restriction in Section 3.6. In settings when k_n is large, which is the focus of this paper, it is possible to allow the instruments to have direct effects on the outcome without losing identification. In Section 3.5, I consider one such relaxation.

Throughout the paper, I assume that the errors are mutually independent and conditionally homoscedastic:

$$\mathbb{E}\left[\begin{pmatrix}\epsilon_i\\ v_{2i}\end{pmatrix} \,\middle|\, W, Z\right] = 0, \qquad \mathbb{E}\left[\begin{pmatrix}\epsilon_i\\ v_{2i}\end{pmatrix}\begin{pmatrix}\epsilon_i\\ v_{2i}\end{pmatrix}' \,\middle|\, W, Z\right] = \Sigma. \tag{3.2}$$

The consistency results in this paper will rely on the additional structure on second moments of the data that the conditional homoscedasticity provides. If heteroscedasticity is a big concern, recent papers by Hausman *et al.* (2012), Chao *et al.* (2012) and Bekker and Crudu (2012) propose to use jackknife type estimators that are consistent under many instruments even in the presence of heteroscedasticity. Those estimators are, however, less efficient under homoscedasticity than the estimators considered here.

In order to employ sufficiency and invariance arguments, I will also assume that the errors are Normally distributed:

Assumption N (Normality). $(\epsilon_i, v_{2i})' \mid W, Z \sim \mathcal{N}(0, \Sigma)$.

This assumption has no effect on consistency results. Normality, however, does have an effect on asymptotic distributions and asymptotic efficiency properties of estimators. I relax this assumption in Section 3.4 when I discuss a minimum-distance approach to inference in this model.

Rather than working with the original regressors W and Z, it will be convenient to work with the orthogonal pair $Z_\perp$ and W, as in Moreira (2003), where the subscript $_\perp$ is a shorthand for taking residuals after regression on the exogenous regressors W, so that for any $n \times d$ matrix A, $A_\perp = (I - W(W'W)^{-1}W')A$. Then the reduced form of the model (3.1) can be written in matrix form as:

$$Y = \begin{pmatrix} Z_\perp & W \end{pmatrix} \begin{pmatrix} \pi_{11,n} & \pi_{12,n} \\ \pi_{21,n} & \pi_{22,n} \end{pmatrix} + V, \quad \text{where} \quad \begin{aligned} \pi_{11,n} &= \pi_{12,n}\beta, \\ \pi_{21,n} &= \delta_n + \pi_{22,n}\beta, \\ \pi_{22,n} &= \tilde{\pi}_{22,n} + (W'W)^{-1}W'Z\pi_{12,n}, \end{aligned} \tag{3.3}$$

where $Y = (y, x) \in \mathbb{R}^{n\times 2}$ with rows $Y_i' = (y_i, x_i)$ pools all endogenous variables in the model, and $V = (\epsilon + \beta v_2, v_2) \in \mathbb{R}^{n\times 2}$ with rows $v_i' = (\epsilon_i + \beta v_{2i}, v_{2i})$ pools the reduced-form errors. The variance of v_i is given by Ω, where

$$\Omega = \Gamma\Sigma\Gamma', \qquad \Gamma = \begin{pmatrix} e_1 & a \end{pmatrix}, \qquad a = \begin{pmatrix} \beta \\ 1 \end{pmatrix}. \tag{3.4}$$

Apart from being excluded from the structural equation, the instruments also have to be relevant in the sense that they have to be correlated with the endogenous variable. To measure the strength of identification I follow Chamberlain (2007) and Andrews, Moreira and Stock (2008) and I use:

$$\lambda_n = \pi'_{12,n} Z'_\perp Z_\perp \pi_{12,n} \cdot a' \Omega^{-1} a / n. \tag{3.5}$$

This parameter is related to the concentration parameter of Rothenberg (1984), which is given by $\pi'_{12,n} Z'_\perp Z_\perp \pi_{12,n} / \Sigma_{22}$. Instead of dividing $\pi'_{12,n} Z'_\perp Z_\perp \pi_{12,n}$ by the (2,2) element of the structural covariance matrix Σ, λ_n multiplies it by the (2,2) element of the structural precision matrix Σ^{-1}, which is given by $a' \Omega^{-1} a$. Therefore, if the structural correlation coefficient $\rho = \Sigma_{12} / \sqrt{\Sigma_{11} \Sigma_{22}}$ is zero, the two measures coincide. Otherwise they are proportional to each other: $\pi'_{12,n} Z'_\perp Z_\perp \pi_{12,n} / \Sigma_{22} = n \lambda_n (1 - \rho^2)$.

The goal is to construct inference procedures that work well even if the number of instruments k_n and the number of exogenous regressors ℓ_n is large relative to sample size. To capture the finite-sample behaviour in these settings in asymptotic approximations, I follow Anatolyev (2011) and Kolesár *et al.* (2011) and allow for many-instrument asymptotics with both k_n and ℓ_n potentially growing in proportion to the sample size:

Assumption MI (Many instruments). (i)

$k_n / n = \alpha_k + o(n^{-1/2})$ and $\ell_n / n = \alpha_\ell + o(n^{-1/2})$ for some $\alpha_\ell, \alpha_k \geq 0$ such that $\alpha_k + \alpha_\ell < 1$; **(ii)**

$\{(z_i, w_i, v_i) \in \mathbb{R}^{k_n} \times \mathbb{R}^{\ell_n} \times \mathbb{R}^2 \colon i = 1, \ldots, n; k_n + \ell_n < n\}_{n \geq 1}$ is a triangular array of iid random variables; **(iii)**

$(W, Z_\perp)$ is full column rank with probability one; and **(iv)**

$\lambda_n \to \lambda$ for some $\lambda > 0$.

Assumption MI (i) weakens the many instrument sequence of Bekker (1994) by allowing ℓ_n to grow with the sample size. The motivation for this is twofold. First, often the presence of a large number of instruments is the result of interacting a few basic instruments with many

exogenous covariates (as in , for example Angrist and Krueger, 1991), in which case both ℓ_n and k_n are large. Second, oftentimes the instruments are valid only conditional on a large set of covariates w_i, such as higher-level fixed effects in multilevel sampling. The remaining parts are standard. Part (ii) allows the distribution of the random variables to change with the sample size. To reflect this, I should index the random variables by n. I drop this index for ease of notation, and only use the subscript n for parameters which change with the sample size. MI (iii) normalizes the first-stage regressors to be full rank. Finally, Part (iv) is the many-instruments equivalent of the relevance assumption and ensures identification. It is equivalent to assuming that the Rothemberg concentration parameter grows at the same rate as the sample size. By allowing $\alpha_k = \alpha_\ell = 0$, Assumption MI nests the standard strong instrument asymptotics.

3.2.2 Sufficient statistics and orthogonal parametrization

Under Normality, the set of sufficient statistics is given by the normalized least-squares estimator of the reduced form coefficients,

$$\begin{pmatrix} \hat{\Pi}_1 \\ \hat{\Pi}_2 \end{pmatrix} = \begin{pmatrix} n^{-1/2}(Z_\perp' Z_\perp)^{-1/2} Z_\perp' Y \\ n^{-1/2}(W'W)^{-1/2} W'Y \end{pmatrix} \in \mathbb{R}^{(k_n+\ell_n)\times 2},$$

and an unbiased estimator of the reduced-form covariance matrix Ω based on the residual sum of squares,

$$S = Y_\perp'(I - Z_\perp(Z_\perp' Z_\perp)^{-1} Z_\perp) Y_\perp / (n - k_n - \ell_n) \in \mathbb{R}^{2\times 2}.$$

The significance of the normalization of the least-squares coefficients is that now the rows of $\hat{\Pi}_1$ and $\hat{\Pi}_2$ are mutually independent. Rather than working with the full set of sufficient statistics, I base inference on $\hat{\Pi}_1$ and S only[1] as in Moreira (2003) and Chamberlain and

[1]Formally, this requirement can be justified by requiring invariance to location shifts in $\hat{\Pi}_2$ in the sample space, and invariance to location shifts in $(\pi_{21,n}, \pi_{22,n})$ in the parameter space. Since the goal is to make inferences about β, the loss function will not depend on $(\pi_{21,n}, \pi_{22,n})$, and will therefore also be invariant to this transformation.

Imbens (2004). Since the distribution of $\hat{\Gamma}_2$ is unrestricted, dropping it from the model does not result in loss of information. This step eliminates the potentially high-dimensional nuisance parameters $\pi_{21,n}$ and $\pi_{22,n}$, so that the model parameters are now given by the triplet $(\beta, \pi_{12,n}, \Omega)$.

Next, to help separate the problem of inference about β from that of the nuisance parameters, I reparametrize $\pi_{12,n}$ as

$$\eta_n = n^{-1/2}(Z_\perp' Z_\perp)^{1/2}\pi_{12,n}\sqrt{a'\Omega^{-1}a}.$$

I refer to η_n are the orthogonalized first-stage coefficients. The advantage of the (β, η_n, Ω) parametrization is that the parameter of interest β is information-orthogonal to the nuisance parameters (η_n, Ω) in the sense that the information matrix is block-diagonal.[2] The distribution of the statistics $\hat{\Gamma}_1$ and S is now given by:

$$\text{vec}(\hat{\Gamma}_1) \sim \mathcal{N}_{2k_n}\left(a \otimes \eta_n (a'\Omega^{-1}a)^{-1/2}, n^{-1}\Omega \otimes I_{k_n}\right), \tag{3.6}$$

$$(n - k_n - \ell_n)S \sim \mathcal{W}_2(n - k_n - \ell_n, \Omega), \tag{3.7}$$

with $\hat{\Gamma}_1$ independent of S. It will be useful to define the following functions of the statistics $\hat{\Gamma}_1$ and S:

$$T = \hat{\Gamma}_1' \hat{\Gamma}_1,$$

$$Q_{\mathcal{S}}(\beta, \Omega) = \frac{b'Tb}{b'\Omega b}, \qquad Q_{\mathcal{T}}(\beta, \Omega) = \frac{a'\Omega^{-1}T\Omega^{-1}a}{a'\Omega^{-1}a}, \qquad b = \begin{pmatrix} 1 \\ -\beta \end{pmatrix},$$

$$m_{\min} = \min \text{eig}(S^{-1}T), \qquad m_{\max} = \max \text{eig}(S^{-1}T).$$

The bigger eigenvalue, $m_{\max}$ will help to determine instrument relevance. On the other hand, the smaller eigenvalue $m_{\min}$ plays a key role in testing the exclusion restriction. The functions $Q_{\mathcal{S}}(\beta, \Omega)$ and $Q_{\mathcal{T}}(\beta, \Omega)$ of T will appear in several objective functions. The

[2] See Cox and Reid (1987) for a discussion of the consequences of orthogonal parametrization in problems with nuisance parameters.

properties of $Q_S(\beta, \Omega)$ and $Q_T(\beta, \Omega)$ are discussed in Andrews, Moreira and Stock (2006).[3]

3.2.3 Limited information likelihood

Inference under Assumption MI is complicated by the incidental parameter problem (Neyman and Scott, 1948): if the number of instruments and exogenous regressors grows with the sample size, the dimension of the nuisance parameters η_n also increases to infinity. Therefore, the standard results about optimality of likelihood-based inference do not apply since they require the dimension of the parameter space to remain fixed.

One way to proceed is to ignore the incidental parameter problem and base inference on the full model likelihood anyway. This likelihood, based on the statistics $\hat{\Pi}_1$ and S, is up to a constant of proportionality given by

$$\mathcal{L}_{\text{LI},n}(\beta, \eta_n, \Omega) \propto |\Omega|^{-(n-\ell_n)/2} e^{-\frac{n}{2}\left[\operatorname{tr}(\Omega^{-1} Y_\perp' Y_\perp / n) - 2\eta_n' \frac{\hat{\Pi}_1 \Omega^{-1} a}{(a'\Omega^{-1}a)^{1/2}} + \eta_n'\eta_n\right]}. \tag{3.8}$$

I refer to this likelihood as the limited-information likelihood after a seminal paper by Anderson and Rubin (1949). In the (β, η_n, Ω) parametrization, the information matrix has a block-diagonal structure:

$$\frac{\mathcal{I}_{\text{LI},n}(\beta, \eta_n, \Omega)}{n} = \begin{pmatrix} \frac{\lambda}{a'\Omega^{-1}a \cdot b'\Omega b} & 0 & 0 \\ 0 & I_{k_n} & \frac{\eta_n}{2a'\Omega^{-1}a}(a \otimes a)'\tilde{D} \\ 0 & \tilde{D}'(a \otimes a)\frac{\eta_n'}{2a'\Omega^{-1}a} & \tilde{D}'\left(\frac{1-\ell_n/n}{2}\Omega \otimes \Omega + \frac{\lambda_n}{4(a'\Omega^{-1}a)^2} aa' \otimes aa'\right)\tilde{D} \end{pmatrix}, \tag{3.9}$$

where $\tilde{D} = (\Omega^{-1} \otimes \Omega^{-1})D_2$ and D_2 is the duplication matrix.[4] The derivations of Equations (3.8) and (3.9) are given in Appendix C.3. Anderson and Rubin (1949) show that the maximum likelihood estimator for β, called the limited-information maximum likelihood

[3] My statistics S and T do not correspond to the statistics S and T in Andrews *et al.* (2006)

[4] see Appendix C.1 for properties of this matrix

(LIML) estimator, solves

$$\hat{\beta}_{\text{LIML}} = \underset{\beta}{\text{argmax}}\, Q_{\mathcal{T}}(\beta, S) = \underset{\beta}{\text{argmin}}\, Q_{\mathcal{S}}(\beta, S) = \frac{T_{12} - m_{\min} S_{12}}{T_{22} - m_{\min} S_{22}}. \tag{3.10}$$

It turns out this maximum likelihood estimator is consistent for β under Assumptions ER and MI despite the incidental parameter problem. Its asymptotic distribution under Normality is given by (see Bekker, 1994 and Kolesár *et al.*, 2011 for derivation)

$$\sqrt{n}\left(\hat{\beta}_{\text{LIML}} - \beta\right) \Rightarrow \mathcal{N}_1\left(0, \mathcal{V}_{\text{LIML},N}\right), \tag{3.11}$$

where

$$\mathcal{V}_{\text{LIML},N} = \frac{b'\Omega b \cdot a'\Omega^{-1}a}{\lambda}\left(1 + \frac{\alpha_k(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell}\frac{1}{\lambda}\right). \tag{3.12}$$

Setting $\alpha_k = \alpha_\ell = 0$ reduces the formula to the standard variance formula under strong-instrument asymptotics. The correction factor in parentheses can be substantial even when the ratio of instruments to sample size, α_k, is small if the normalized concentration parameter λ is small. The presence of many exogenous regressors, the case when $\alpha_\ell > 0$, has a negligible impact on the asymptotic variance unless α_ℓ is large. Their presence does, however, have a big impact on tests for overidentifying restrictions as I discuss in Section 3.6. In addition to being consistent, $\hat{\beta}_{\text{LIML}}$ is also asymptotically efficient among the class of estimators invariant to rotations of instruments (see Chioda and Jansson, 2009).

There are, nonetheless, two problems with this approach. First, the limited information likelihood cannot be used for inference, since its curvature is too big— the (1,1) element of the inverse information matrix (3.9), $(\mathcal{I}_{\text{LI}}^{-1})_{11} = \mathcal{I}_{\text{LI},11}^{-1}$ is missing the factor in parentheses that appears in the correct formula (3.12). As a result, confidence intervals for $\hat{\beta}_{\text{LIML}}$ based on the limited information likelihood will undercover. Instead, confidence intervals have to be computed using the correct asymptotic formula (3.11). However, a simple plug-in procedure does not work because the maximum likelihood estimators of λ_n and Ω are inconsistent (see Appendix C.3 for derivation):

$$\hat{\lambda}_{\text{LIML}} = \frac{n-\ell_n}{n-k_n-\ell_n} m_{\max} \xrightarrow{p} \frac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}(\lambda+\alpha_k), \tag{3.13a}$$

$$\hat{\Omega}_{\text{LIML}} = \frac{n-k_n-\ell_n}{n-\ell_n} S + \frac{n}{n-\ell_n}\left(T - \frac{m_{\max}\hat{a}_{\text{LIML}}\hat{a}'_{\text{LIML}}}{\hat{a}'_{\text{LIML}}S^{-1}\hat{a}_{\text{LIML}}}\right) \xrightarrow{p} \Omega - \frac{\alpha_k}{1-\alpha_\ell}\frac{aa'}{a'\Omega^{-1}a}. \tag{3.13b}$$

Bekker (1994) and Hansen *et al.* (2008) therefore modify the simple plug-in procedure using estimators for λ and Ω that are consistent under MI when $\alpha_\ell = 0$. These asymptotic variance estimators, however, have to be modified again if we want to allow $\alpha_\ell > 0$ (Anatolyev, 2011; Kolesár *et al.*, 2011).

The second problem is that it is unclear how to modify the likelihood (3.8) so that it delivers a consistent estimator when Assumption ER is relaxed to allow the instruments to have direct effects on the outcome.

In the next section, I introduce an alternative (quasi-) likelihood approach that addresses both of these problems.

3.3 Equivalence between Integrated and Random Effects Likelihoods

This section derives the basic result of the paper that we can use an invariance argument and the Bernstein-von Mieses theorem to construct an integrated likelihood that addresses the incidental parameter problem.

The idea behind using an invariance argument is that if we require inference to be invariant to suitably chosen group actions, the maximal invariant in the parameter space will preserve β, and it will have a fixed dimension even as the number of instruments grows. I follow Andrews *et al.* (2006), Chamberlain (2007), Chioda and Jansson (2009), and Moreira (2009), and I consider transformations given by

$$\overline{m}_1(g,(\hat{\Pi}_1,S)) = (g\hat{\Pi}_1,S), \qquad \overline{m}_2(g,(\beta,\eta_n,\Omega)) = (\beta,g\eta_n,\Omega), \qquad g\in\mathcal{O}(k_n),$$

where $\mathcal{O}(k_n)$ is the group of $k_n \times k_n$ orthogonal matrices. Here $\overline{m}_1$ is the action on the sample

space, and it rotates the direction of the instruments. Correspondingly, $\overline{m}_2$, the action on the parameter space, rotates the direction of the first-stage coefficients η_n, but not their norm $\lambda_n^{1/2} = \sqrt{\eta_n'\eta_n}$, where λ_n is the scalar measure of instrument strength defined in Equation (3.5). Invariant decision rules will therefore not depend on the direction of the instruments. It is straightforward to show that the maximal invariants are given by $T = \hat{\Pi}_1'\hat{\Pi}_1$ and S on the sample space, and $(\beta, \lambda_n, \Omega)$ on the parameter space. The potentially high-dimensional vector of first-stage coefficients η_n has been reduced to a scalar.

The parameter space of the maximal invariant (S, T) is given by the maximal invariant on the parameter space, $(\beta, \lambda_n, \Omega)$, which has a fixed dimension irrespective of the number of instruments. Since the likelihood based on the maximal invariant, $\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T)$, which I call the invariant likelihood, is smooth, it is locally asymptotically Normal under many-instrument asymptotics (Chioda and Jansson, 2009). Therefore, inference based on the invariant likelihood will be asymptotically efficient among invariant procedures by standard arguments (see, for example, van der Vaart, 1998, Chapter 8). Moreira (2009) shows that if Ω is known, the maximum invariant likelihood estimator for β coincides with LIMLK, which is indeed asymptotically efficient among invariant estimators. The next proposition generalizes this result:

Proposition 3.1. *The* MLE *based on the invariant likelihood* $\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T)$ *is given by* $\hat{\beta}_{\text{LIML}}$. *This result also holds if* λ_n *is fixed at an arbitrary value.*

The first part of the proposition generalizes Moreira's result to the case when Ω is not known, and shows that the maximal invariant likelihood estimator then coincides with LIML. Since LIML is efficient among regular invariant estimators, this result confirms that maximizing the invariant likelihood indeed produces an efficient invariant estimator. Furthermore, this result also explains why the limited-information likelihood produces an estimator that is robust to many instruments: it is because LIML happens to coincide with the maximum invariant likelihood estimator.

The second part of the proposition shows that constraining λ_n to be equal to a particular value does not affect the maximum invariant likelihood estimate. This result is similar to

that in Chamberlain (2007) who shows that the Bayes rule under a particular loss function and prior for β does not depend on the prior for λ_n. Since the information matrix of the invariant likelihood is block-diagonal between λ_n and β, the maximum likelihood estimate of β when λ_n is given should vary only slowly with λ_n (see Cox and Reid, 1987, Section 2.2). The proposition shows that the dependence is even more limited: the estimate does not vary with λ_n at all.

There is an alternative way of building the invariant likelihood that will allow me to use this result to build a connection between it and the random-effects likelihood. The argument is similar to that in Chamberlain and Moreira (2009), who relate invariant likelihood to a correlated random effects likelihood in a dynamic panel data model. In particular, imposing invariance is equivalent to assuming a particular prior distribution for the model parameters, induced by the Haar measure on $\mathcal{O}(k_n)$, called the invariant prior distribution (Eaton, 1989). Since the group $\mathcal{O}(k_n)$ is compact, this prior is unique. Consider a polar decomposition of the first stage coefficients:

$$\eta_n = \omega_n \lambda_n^{1/2}, \qquad \omega_n = \eta_n / \|\eta_n\|, \qquad \lambda_n = \|\eta_n\|^2.$$

The potentially high-dimensional nuisance parameter ω_n is a point on the unit sphere that measures the direction of η_n. Under this decomposition, the invariant prior is given by the uniform distribution over the unit sphere $\mathbb{S}^{k_n-1}$ in $\mathbb{R}^{k_n}$, the parameter space for the parameter ω_n. Furthermore, the invariant likelihood is equivalent to the integrated (marginal) likelihood that uses this invariant prior as a prior distribution. Denoting the prior by $F_{\omega_n}(\cdot)$, this relationship can be written as

$$\mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T) = \int_{\mathbb{S}^{k_n-1}} \mathcal{L}_{\text{LI},n}(\beta, \lambda_n, \omega_n, \Omega; \hat{\Pi}_1, S) \, \mathrm{d}F_{\omega_n}(\omega_n), \tag{3.14}$$

where $\mathcal{L}_{\text{LI},n}$ is the limited information likelihood given in Equation (3.8).

One disadvantage of the invariant likelihood is that due to the presence of Bessel functions in the likelihood expression, estimates of λ_n and Ω are not available in closed form and have to be computed by maximizing the invariant likelihood numerically. This makes

construction of likelihood-based confidence intervals for β difficult, since these estimates are needed for evaluating the Hessian. Therefore, although the inverse Hessian evaluated at maximum likelihood estimates is a consistent estimator of the asymptotic variance of $\hat{\beta}_{\text{LIML}}$, getting Hessian-based standard error estimates involves numerical optimization.

This motivates an introduction of a prior over λ_n, in addition to the uniform prior over ω_n. If this additional prior is appropriately chosen, integrating the limited information likelihood over both priors will yield an integrated likelihood that is more convenient to work with than the invariant likelihood. Since by Proposition 3.2, constraining λ_n does not affect the maximum invariant likelihood estimator for β, introducing a prior for λ_n will not affect it either: it will still be given by $\hat{\beta}_{\text{LIML}}$. Moreover, so long as this low-dimensional prior is not dogmatic, the Bernstein von-Mises theorem should apply, and the prior should get dominated in large samples. Therefore inference based on the integrated likelihood should agree with inference based on the invariant likelihood in large samples.

The family of priors I consider is a scaled chi-square family with an unknown scale parameter $\lambda > 0$:

$$\lambda_n \sim \frac{\lambda}{k_n}\chi^2(k_n). \tag{3.15}$$

The hyperparameter λ in this prior corresponds to the limit of λ_n under Assumption MI. I allow it to be determined by the data, so that the prior will be dominated in large samples. This prior and the uniform prior over ω_n are equivalent to a single Normal prior over η,

$$\eta_n \sim \mathcal{N}(0, \lambda/k_n), \tag{3.16}$$

which corresponds to the random-effects prior proposed in Chamberlain and Imbens (2004). Therefore, the integrated likelihood obtained after integrating the limited information likelihood in Equation (3.8) over the invariant prior on ω_n and the chi-square prior on λ_n coincides with the RE likelihood that integrates the limited information likelihood over a single Normal prior (3.16). The RE likelihood, unlike the invariant likelihood, has a simple

closed form (see Appendix C.3 for derivation):

$$\begin{aligned}\mathcal{L}_{\text{RE},n}(\beta,\lambda,\Omega) &= \int_{\mathbb{R}^{k_n}} \mathcal{L}_{\text{INV},n}(\beta,\eta_n,\omega_n,\Omega;T,S)\,\mathrm{d}F_{\eta_n|\lambda}(\eta_n \mid \lambda) \\ &= \int_{\mathbb{R}}\int_{\mathbb{S}^{k_n-1}} \mathcal{L}_{\text{LI},n}(\beta,\lambda_n,\omega_n,\Omega;\hat{\Pi}_1,S)\,\mathrm{d}F_{\omega_n}(\omega_n)\,\mathrm{d}F_{\lambda_n|\lambda}(\lambda_n \mid \lambda) \\ &= \left(1+\frac{n}{k_n}\lambda\right)^{-k_n/2} |\Omega|^{-(n-\ell_n)/2} e^{-\frac{1}{2}\operatorname{tr}(\Omega^{-1}((n-k_n-\ell_n)S+nT))+\frac{n}{2}\frac{\lambda}{n/k_n+\lambda}Q_{\mathcal{T}}(\beta,\Omega)}.\end{aligned} \tag{3.17}$$

This equivalence shows that there are two ways of thinking about the RE assumption (3.16) that the first-stage coefficients η_n are Normally distributed with zero mean and unknown variance. The first is to view it as a modelling tool that reduces the original high-dimensional model to a model in which the parameter space stays 5-dimensional even as $\ell_n \to \infty$ and $k_n \to \infty$. This model is locally asymptotically Normal. Therefore, if the RE assumption holds, inference based on the RE likelihood will have the usual asymptotic optimality properties—maximum likelihood estimators, Wald, LM and LR test will be asymptotically efficient, and the inverse Hessian will be a consistent estimator for the asymptotic variance.

The second way of thinking about the RE assumptions is to view it as arising from two priors. The uniform prior over ω_n can be motivated by invariance arguments. Moreover, Chamberlain (2007) shows this prior is least favourable, so that it can also be motivated by finite-sample minimax considerations. The prior on λ_n is used to make inference more convenient and will not matter asymptotically. Therefore, asymptotic validity and optimality properties of inference based on the RE likelihood are preserved even if we drop the RE assumption, and do not require that the orthogonalized first-stage coefficients are Normally distributed.

Proposition 3.2. *Consider the model* (3.1)–(3.2)*.*

(i) *Suppose that* $m_{\max} > k_n/n$*. Then the maximum likelihood estimators based on the* RE

likelihood (3.17) *are given by:*

$$\begin{aligned}\hat{\beta}_{\text{RE}} &= \hat{\beta}_{\text{LIML}},\\ \hat{\lambda}_{\text{RE}} &= m_{\max} - k_n/n,\\ \hat{\Omega}_{\text{RE}} &= \frac{n-k_n-\ell_n}{n-\ell_n}S + \frac{n}{n-\ell_n}\left(T - \frac{\hat{\lambda}_{\text{RE}}}{\hat{a}_{\text{RE}}' S^{-1}\hat{a}_{\text{RE}}}\hat{a}_{\text{RE}}\hat{a}_{\text{RE}}'\right).\end{aligned}$$

(ii) *Under Assumptions ER, N, and MI,* $(\hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}}) \xrightarrow{p} (\lambda, \Omega)$.

Part (i) of Proposition 3.2 formalizes the claim that the estimator of β remains unchanged under the additional chi-square prior for λ_n. Part (ii) of Proposition 3.2 shows that, unlike estimators based on the limited information likelihood given in Equation (3.13), the RE estimators of λ and Ω are consistent under many instrument asymptotics. The assumption that $m_{\max} \geq k_n/n$ makes sure that the constraint $\lambda \geq 0$ does not bind when maximizing the likelihood. It will hold in large samples if Assumption MI (iv) holds.

Proposition 3.3. *Consider the model* (3.1)–(3.2).

(i) *The (1,1) element of the inverse Hessian of the random-effects likelihood* (3.17)*, evaluated at* $(\hat{\beta}_{\text{RE}}, \hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$*, is given by:*

$$\hat{\mathcal{H}}^{11}_{\text{RE}} = \frac{\hat{b}_{\text{RE}}'\hat{\Omega}_{\text{RE}}\hat{b}_{\text{RE}}(\hat{\lambda}_{\text{RE}} + k_n/n)}{n\hat{\lambda}_{\text{RE}}}\left(\hat{Q}_{\mathcal{S}}\hat{\Omega}_{\text{RE},22} - T_{22} + \frac{\hat{c}}{1-\hat{c}}\frac{\hat{Q}_{\mathcal{S}}}{\hat{a}_{\text{RE}}'\hat{\Omega}_{\text{RE}}^{-1}\hat{a}_{\text{RE}}}\right)^{-1},$$

where $\hat{Q}_{\mathcal{S}} = Q_{\mathcal{S}}(\hat{\beta}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$ *and* $\hat{c} = \frac{\hat{\lambda}_{\text{RE}}\hat{Q}_{\mathcal{S}}}{(k_n/n+\hat{\lambda}_{\text{RE}})(1-\ell_n/n)}$.

(ii) *Under Assumptions ER, N, and MI,* $-n\hat{\mathcal{H}}^{11}_{\text{RE}} \xrightarrow{p} \mathcal{V}_{\text{LIML},N}$*, where* $\mathcal{V}_{\text{LIML},N}$ *is given in Equation* (3.12).

This result proves that the extra prior on λ_n gets dominated in large samples so that the inverse Hessian can be used to estimate the asymptotic variance of $\hat{\beta}_{\text{RE}}$.

The key condition for Proposition 3.3 to hold is that the extra prior on λ_n is not dogmatic. For example, Lancaster (2002) suggests to deal with incidental parameters in panel data models by first orthogonalizing them, and then integrating them out with respect to a suitable uniform prior. In the instrumental variables model the parameter space for the

orthogonalized parameters η_n is $\mathbb{R}^k$, so that a "uniform prior" corresponds to a flat prior on $\mathbb{R}^k$, which in turn corresponds to a uniform prior on ω_n, and an improper prior on λ_n, obtained by taking the limit as $\lambda \to \infty$ of the chi-square prior (3.15). The integrated likelihood based on this prior corresponds to the limit of the RE likelihood (3.17) as $\lambda \to \infty$:

$$\lim_{\lambda \to \infty} \mathcal{L}_{\text{RE},n}(\beta, \lambda, \Omega) = |\Omega|^{-(n-\ell_n)/2} e^{-\frac{1}{2}\operatorname{tr}(\Omega^{-1}((n-k_n-\ell_n)S+nT))+\frac{n}{2}Q_T(\beta,\Omega)}.$$

This objective function coincides with the concentrated limited information likelihood that concentrates η_n out, and therefore does not produce valid confidence intervals, since the prior on λ puts all its mass far away from zero. On the other hand, this dogmatic prior on λ_n does not affect the consistency of the maximum integrated likelihood estimator of β as the second part of Proposition 3.1 predicts.

3.4 Efficient minimum distance estimation under non-Normal errors

Identification in the invariant model comes from restrictions on the expectation of the invariant statistics S and T imposed by the exclusion restriction. The Normality assumption on the errors plays no role. This observation motivates a minimum distance objective function. In this section, I first show that the random effects estimator is in fact equivalent to a minimum distance estimator that uses a particular weight matrix. This weight matrix weights the restrictions efficiently under Normality, but not otherwise. Second, I derive an efficient minimum distance estimator when the Normality assumption is dropped, and use the equivalence result to construct minimum-distance based standard errors for LIML that are valid under non-Normality.

To simplify the expressions in this section, let D_2 denote the duplication matrix, L_2 the elimination matrix and N_2 the symmetrizer matrix. The duplication matrix transforms the vech operator into a vec operator[5], and the elimination operator performs the reverse

[5]The operator $\operatorname{vec}(A)$ stacks columns of A into a single column. The operator $\operatorname{vech}(A)$ transforms the

operation, so that $D_d \operatorname{vech}(A) = \operatorname{vec}(A)$, and $L_d \operatorname{vec}(A) = \operatorname{vech}(A)$, where $A \in \mathbb{R}^{d\times d}$. The symmetrizer matrix has the property that $N_d \operatorname{vec}(A) = (1/2)\operatorname{vec}(A + A')$. Other properties of these matrices are given in Appendix C.1.

3.4.1 Random effects and minimum distance

The instrumental variables model (3.1)–(3.2) without any further assumptions implies that

$$\mathbb{E}[S] = \Omega, \quad \mathbb{E}[T] = (k_n/n)\Omega + \Xi_n, \quad \Xi_n = \frac{1}{n}\begin{pmatrix} \pi_{11,n}{}'Z_\perp' Z_\perp \pi_{11,n} & \pi_{11,n}{}'Z_\perp' Z_\perp \pi_{12,n} \\ \pi_{12,n}{}'Z_\perp' Z_\perp \pi_{11,n} & \pi_{12,n}{}'Z_\perp' Z_\perp \pi_{12,n} \end{pmatrix}. \tag{3.18}$$

Since the parameters Ω, $\pi_{11,n}$ and $\pi_{12,n}$ are unrestricted, these two expectations are unrestricted. Under assumption ER, however, the second-stage coefficients $\pi_{11,n}$ are restricted to be proportional to the first stage coefficients: $\pi_{11,n} = \pi_{12,n}\beta$. This restriction leads to a rank restriction on Ξ_n, namely that $\Xi_n = (a'\Omega^{-1}a)^{-1}\lambda_n aa'$. This rank restriction can be used to build a minimum distance (MD) objective function

$$\mathcal{Q}_n(\beta, \lambda_n, \Omega; \hat{W}_n) = \begin{pmatrix} \operatorname{vech}(S - \Omega) \\ \operatorname{vech}\left(T - (k_n/n)\Omega - \frac{\lambda_n aa'}{a\Omega^{-1}a}\right) \end{pmatrix}' \hat{W}_n \begin{pmatrix} \operatorname{vech}(S - \Omega) \\ \operatorname{vech}\left(T - (k_n/n)\Omega - \frac{\lambda_n aa'}{a\Omega^{-1}a}\right), \end{pmatrix} \tag{3.19}$$

where $\hat{W}_n \in \mathbb{R}^{6\times 6}$ is a weight matrix. In the random effects model, the identification of the model coefficients is based on the same restriction. The only difference is that the parameter λ_n is replaced by its expectation under the chi-square prior, λ. There should therefore exist a weight matrix such that the random effects estimator of (β, λ, Ω) is asymptotically equivalent to a minimum distance estimator with respect to this weight matrix. The next proposition shows that if the weight matrix is chosen carefully, the minimum distance and random effects estimators are in fact *identical*.

---

lower-triangular part of A into a single column.

Proposition 3.4. *Suppose that* $\operatorname{tr}(S^{-1}T) > 2k_n/n$. *Then the minimum distance estimator based on minimizing the objective function* (3.19) *with respect to the weight matrix*

$$\hat{W}_{\text{RE}} = \begin{pmatrix} (n - k_n - \ell_n)/n & 0 \\ 0 & (k_n/n)^{-1} \end{pmatrix} \otimes (D_2'(S^{-1} \otimes S^{-1})D_2) \tag{3.20}$$

is given by $(\hat{\beta}_{\text{RE}}, \hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$

If the errors are Normally distributed, then the weight matrix $\hat{W}_{\text{RE}}$ weights the moment conditions (3.18) efficiently under many-instrument asymptotics, even though it is not proportional to the inverse of the asymptotic covariance matrix of $(\operatorname{vech}(S)', \operatorname{vech}(T)')'$—the condition that the weight matrix converges to the inverse of the asymptotic covariance matrix of the moment conditions is sufficient, but not necessary for asymptotic efficiency (Newey and McFadden, 1994, Section 5.2).

Proposition 3.5. *Consider the model* (3.1)–(3.2), *and suppose that Assumptions ER, N, and MI hold. Consider a minimum distance estimator based on the objective function* (3.19). *Suppose that, for some constant* $c \in \mathbb{R}$, *the weight matrix satisfies*

$$\hat{W}_n \xrightarrow{p} \begin{pmatrix} (1 - \alpha_k - \alpha_\ell)D_2'\Omega^{-1} \otimes \Omega^{-1}D_2 & 0 \\ 0 & D_2' \left[\alpha_k \Omega \otimes \Omega + c\Omega \otimes (aa') + c(aa') \otimes \Omega\right]^{-1} D_2. \end{pmatrix}$$

Then the minimum distance estimator for (β, λ, Ω) *is optimal among the class of minimum distance estimators.*

If $c = \lambda/(a'\Omega^{-1}a)$, then the limit weight in Proposition 3.5 equals twice the inverse of the asymptotic covariance matrix of $(\operatorname{vech}(S)', \operatorname{vech}(T)')'$. The proposition shows that it is possible to misspecify λ in the optimal weight without affecting the asymptotic distribution of the minimum distance estimator. In particular, the weight matrix $\hat{W}_{\text{RE}}$ satisfies the condition in Proposition 3.5 with $c = 0$. The random effects estimator, therefore, can also be viewed as an efficient minimum distance estimator.

This result is similar to Goldberger and Olkin (1971), who consider a minimum distance objective function based on the proportionality restriction that the exclusion restriction

imposes on the expectation of $\hat{\Pi}_1$:

$$\mathcal{Q}_{\text{GO},n}(\beta, \pi^*_{12,n}) = \text{vec}\left(\hat{\Pi}_1 - \pi^*_{12,n}a'\right)' \left(S^{-1} \otimes I_{k_n}\right) \text{vec}\left(\hat{\Pi}_1 - \pi^*_{12,n}a'\right), \tag{3.21}$$

where $\pi^*_{12,n} = (Z'_\perp Z_\perp/n)^{1/2}\pi_{12,n}$. Goldberger and Olkin (1971) show that this objective function is minimized at $\hat{\beta}_{\text{LIML}}$. The weight matrix $S^{-1} \otimes I_{k_n}$ consistently estimates the inverse of the asymptotic variance of $\text{vec}(\hat{\Pi}_1)$ under strong-instrument asymptotics.

The efficiency result in Proposition 3.5 is, however, sensitive to the assumption of Normality. If this assumption is dropped, the asymptotic covariance matrix of $(\text{vech}(S)', \text{vech}(T)')'$ loses its special structure, and the minimum distance estimator based on the efficient weight matrix will in general have lower asymptotic variance than $\hat{\beta}_{\text{LIML}}$. This sensitivity to the assumption of Normality is similar to the result in panel-data models where identification is based on covariance restrictions; there the weight matrix used by the maximum likelihood estimator is only optimal under Normality (Arellano, 2003, Chapter 5.4).

3.4.2 Efficient minimum distance estimator under non-Normal errors

The covariance matrix of the joint asymptotic distribution of S and T when the errors are not Normal is quite complicated. Therefore, in order to simplify the derivation of the efficient MD estimator, I will work with a one-to-one transformation of the moment conditions that will allow me to reduce their dimension. Let $\Lambda_{22,n} = \lambda_n/(a'\Omega^{-1}a)$, and $\Lambda_{22} = \lambda/(a'\Omega^{-1}a)$, and consider a one-to-one transformation of the moment conditions (3.18) given by

$$\mathbb{E}[S] = \Omega, \qquad\qquad \mathbb{E}[T - (k_n/n)S] = \Lambda_{22,n}aa'.$$

The parameter space is now given by $(\beta, \Lambda_{22,n}, \Omega)$, and the nuisance parameter Ω only appears in the first moment condition. Since the first moment condition is unrestricted, minimizing an objective function that only uses the second moment condition with respect to an efficient weight will yield an estimator of β that has the same asymptotic variance as the efficient minimum distance that uses both of them (Chamberlain, 1982, Section 3.2).

Therefore, the objective function (3.19) can be simplified as

$$\mathcal{Q}_{\text{SIMP},n}(\beta, \Lambda_{22}; \hat{W}_n) = \text{vech}\left(T - (k_n/n)S - \Lambda_{22,n}aa'\right)' \hat{W}_n \, \text{vech}\left(T - (k_n/n)S - \Lambda_{22,n}aa'\right). \tag{3.22}$$

Now the distribution of the moment conditions only depends on a three-dimensional statistic $\text{vech}(T - (k_n/n)S)$, which can be written as a quadratic form:

$$T - (k_n/n)S = Y_\perp' H Y_\perp = (Z_\perp \pi_{12,n} a' + V)' H (Z_\perp \pi_{12,n} a' + V),$$

where

$$H = \frac{1}{n - k_n - \ell_n}\left[(1 - \ell_n/n) Z_\perp (Z_\perp' Z_\perp) Z_\perp' - (k_n/n)(I_n - W(W'W)^{-1}W')\right].$$

In order to derive the limiting distribution of the moment conditions under non-normality, I need to impose some regularity conditions on the matrix of the quadratic form H as well as $Z_\perp \pi_{12,n} a'$, the vector of means of $Y_\perp$:

Assumption RC (Regularity conditions). (i)

The reduced-form errors v_i are iid with finite fourth moments; **(ii)**

For some $\delta, \mu \in \mathbb{R}$, $d'd/n \to \delta$ and $n^{-1}\pi_{12,n}' Z_\perp' d \to \mu$ where $d = \text{diag}(H) \in \mathbb{R}^n$; and **(iii)**

For some constant $C \in \mathbb{R}$, $\sup_i \|(Z_\perp)_i' \pi_{12,n}\| < C < \infty$ and $\sup_n \sup_{1 \le i \le n} \sum_{s=1}^n |(H)_{si}| < C < \infty$.

Part (i) relaxes the Normality assumption on the errors. Part (ii) ensures that all terms in the asymptotic covariance matrix are well-defined. Part (iii) implies that a Lindeberg-type condition holds.

Lemma 3.1. *Consider the model* (3.1)–(3.2). *Then, under Assumptions ER, MI, and RC:*

(i)

$$\sqrt{n}\,\text{vech}(T - (k_n/n)S - \Lambda_{22,n}aa') \Rightarrow \mathcal{N}(0, \Delta), \qquad \Delta = 2L_2 N_2 (\Delta_1 + \Delta_2 + \Delta_3 + \Delta_3') L_2',$$

where

$$\Delta_1 = \Lambda_{22} aa' \otimes \Omega + \Omega \otimes \Lambda_{22} aa' + \tau \Omega \otimes \Omega, \qquad \tau = \alpha_k (1 - \alpha_\ell)/(1 - \alpha_k - \alpha_\ell),$$
$$\Delta_2 = \frac{\delta}{2}\left[\Psi_4 - \operatorname{vec}(\Omega)\operatorname{vec}(\Omega)' - 2\Omega \otimes \Omega\right], \qquad \Psi_4 = \mathbb{E}[(v_i v_i') \otimes (v_i v_i')],$$
$$\Delta_3 = \mu \Psi_3' \otimes a, \qquad \Psi_3 = \mathbb{E}[(v_i v_i') \otimes v_i].$$

(ii) *Let $M = I_n - Z_\perp (Z_\perp' Z_\perp) Z_\perp' - W(W'W)^{-1}W'$, and let $\hat{V} = MY$ with rows $\hat{v}_i$ denote estimates of the reduced-form errors. If the errors v_i have finite eighth moments, then*

$$\hat{\Psi}_3 = \frac{\sum_i [(\hat{v}_i \hat{v}_i') \otimes \hat{v}_i]}{\sum_i M_{ii}^3} \xrightarrow{p} \Psi_3,$$
$$\hat{\Psi}_4 = \frac{\sum_i (\hat{v}_i \hat{v}_i') \otimes (\hat{v}_i \hat{v}_i') - \left[\delta_M - \sum_{ij} M_{ij}^4\right](2N_2 \hat{\Omega} \otimes \hat{\Omega} + \operatorname{vec}(\hat{\Omega})\operatorname{vec}(\hat{\Omega})')}{\sum_{ij} M_{ij}^4} \xrightarrow{p} \Psi_4,$$

where $\delta_M = \sum_i M_{ii}^2$.

Part (i) shows that the asymptotic variance consists of three distinct terms. If the errors are Normally distributed, then $\Delta_2 = \Delta_3 = 0$. The term Δ_2 accounts for excess kurtosis of the errors, and the term Δ_3 accounts for skewness. Part (ii) provides consistent estimators for the third and fourth moments of the errors. Since the probability limits of S and T do not depend on Assumption N, the other components of Δ_1, Δ_2 and Δ_3 can be consistently estimated by $\hat{\beta}_{\text{RE}}$, $\hat{\Omega}_{\text{RE}}$, and $\hat{\Lambda}_{22,\text{RE}} = \hat{\lambda}_{\text{RE}}/(\hat{a}_{\text{RE}}' \hat{\Omega}_{\text{RE}}^{-1} \hat{a}_{\text{RE}})$. Therefore, a consistent estimator of the asymptotic covariance matrix Δ is given by

$$\hat{\Delta} = 2L_2 N_2 (\hat{\Delta}_1 + \hat{\Delta}_2 + \hat{\Delta}_3 + \hat{\Delta}_3') L_2',$$

where the terms $\hat{\Delta}_j$ are given by replacing β, Λ_{22}, and Ω in the definitions of Δ_1, Δ_2 and Δ_3 by their random-effects estimators, and replacing Ψ_3 and Ψ_4 by $\hat{\Psi}_3$ and $\hat{\Psi}_4$. Using this weight in the minimum distance objective function (3.22) yields an efficient minimum distance (EMD) estimator

$$(\hat{\beta}_{\text{EMD}}, \hat{\Lambda}_{22,\text{EMD}}) = \underset{\beta, \Lambda_{22}}{\operatorname{argmin}}\, \mathcal{Q}_{\text{SIMP},n}(\beta, \Lambda_{22}; \hat{\Delta}^{-1}).$$

Since the objective function is a fourth-order polynomial in two arguments, the solution can be easily found numerically. It then follows by standard arguments (see, for example, Newey and McFadden, 1994), that

$$\sqrt{n}(\hat{\beta}_{\text{EMD}} - \beta) \Rightarrow \mathcal{N}(0, \mathcal{V}_{\text{EMD}}),$$

where $\mathcal{V}_{\text{EMD}}$ is given by the (1,1) element of the matrix $(G'\Delta^{-1}G)^{-1}$, where G is the derivative of the moment condition

$$G = L_2 \begin{pmatrix} \Lambda_{22}(a \otimes e_1 + e_1 \otimes a) & a \otimes a \end{pmatrix}.$$

A consistent plug-in estimator of $\mathcal{V}_{\text{EMD}}$ can be easily constructed by replacing Δ by $\hat{\Delta}$, and replacing Λ_{22} and β in the expression for G by their random-effects, or EMD estimators.

The simplified objective function (3.22) is also useful for deriving standard errors for LIML that are consistent under non-Normality since the random effects estimator also minimizes the simplified objective function with respect to a particular weight matrix:

Lemma 3.2. *Suppose that* $\text{tr}(S^{-1}T) \geq 2k_n/n$. *Then the minimum distance estimator based on the objective function* (3.22) *with respect to the weight matrix* $\hat{W}_{\text{SIMP,RE}} = D_2'(S^{-1} \otimes S^{-1})D_2$ *is given by* $(\hat{\beta}_{\text{RE}}, \hat{\Lambda}_{22,\text{RE}})$.

The asymptotic variance of $\hat{\beta}_{\text{LIML}}$ under Assumptions ER, MI, and RC is therefore given by the (1,1) element of the matrix

$$(G'WG)^{-1}G'W\Delta WG(G'WG)^{-1}, \tag{3.23}$$

where $W = D_2'(\Omega^{-1} \otimes \Omega^{-1})D_2 = \text{plim}\, \hat{W}_{\text{SIMP,RE}}$. This element evaluates as

$$\mathcal{V}_{\text{LIML}} = \mathcal{V}_{\text{LIML},N} + \frac{2\mu}{\Lambda_{22}^2}\mathbb{E}[v_{\backslash\epsilon}\epsilon^2] + \frac{\delta}{\Lambda_{22}^2}\mathbb{E}[\epsilon^2 v_{\backslash\epsilon}^2 - |\Omega|],$$

where $v_{\backslash\epsilon} = v_2 - \Sigma_{12}\Sigma_{11}^{-1}\epsilon$ is the part of the first-stage error that is uncorrelated with ϵ, the error in the structural equation. The term $\mathcal{V}_{\text{LIML},N}$ (given in Equation (3.12)) corresponds to

the asymptotic variance of $\hat{\beta}_{\text{LIML}}$ under Normal errors. The two remaining terms are corrections for skewness and excessive kurtosis. Anatolyev (2011) derives the same asymptotic variance expression by working with the explicit definition of LIML. If $\alpha_\ell = 0$, then $\mathcal{V}_{\text{LIML}}$ reduces to the asymptotic variance given in Hansen *et al.* (2008), Anderson *et al.* (2010), and van Hasselt (2010).

A consistent plug-in estimator of this variance can easily be computed by replacing Δ by $\hat{\Delta}$ and replacing a and Ω in the expressions for G and W by $\hat{a}_{\text{RE}}$ and $\hat{\Omega}_{\text{RE}}$, and plugging the estimates $\hat{G}$, $\hat{W}$, and $\hat{\Omega}$ into the expression (3.23).

In general, the optimal minimum distance estimator will be more efficient than LIML. However, apart from the case when the errors are Normal, there are two other cases when the variances are equal. First, Anderson *et al.* (2010) show that when the errors belong to the family of elliptically contoured distributions, then LIML is efficient in the class of estimators that depend on the data only through smooth functions of S and T. Since the efficient minimum distance estimator is a member of this class, the equality $\mathcal{V}_{\text{LIML}} = \mathcal{V}_{\text{EMD}}$ also holds in this case. Second, when $\delta = 0$ (which by the Cauchy-Schwarz inequality implies $\mu = 0$), the terms Δ_2 and Δ_3 in the asymptotic variance of the moment condition drop out, and the result in Proposition 3.5 that the optimal weight matrix can be misspecified again applies.

3.5 Allowing for direct effects of instruments on outcome

In many applied problems, Assumption ER is too restrictive. For example, when the vector of instruments consists of group indicators, it rules out any clustering at the group level in the structural equation. In this section, I consider an approach to inference that is robust to such group-level clustering, so long as the cluster effects are uncorrelated with the effects of the instrument on the endogenous variable. I first discuss the motivation for this weaker identifying assumption. Then I generalize the random effects framework to accommodate such direct effects.

3.5.1 The direct effects problem

To explain the motivation behind relaxing Assumption ER, consider an example from Chetty *et al.* (2011). Chetty *et al.* (2011) are interested in estimating the effect of early childhood achievement, as measured by kindergarten test scores, on subsequent outcomes. For concreteness, take the outcome of interest to be first-grade scores. In the STAR experiment, children and teachers were randomly assigned to kindergarten classrooms, generating an exogenous variation in kindergarten test scores. Assuming that teachers only affect subsequent outcomes through their effect on test scores, we should therefore be able to use kindergarten teacher indicators as instruments for kindergarten test scores. The problem is, however, that since classes mostly stay together in subsequent years, the instrument also affects outcomes directly: the kindergarten teacher indicator coincides with kindergarten classroom indicator, which also has an effect on outcomes through the first-grade teacher. We cannot partial out the effect of first-grade teachers since their assignment is perfectly correlated with kindergarten teacher assignment. More generally, allowing for direct effects of this type is important in applications in which the instrument is only assigned at a group level, but there is a concern that there are other factors, which also vary at a group level, that influence outcomes.

If such direct effects are present, the unrestricted reduced form of the model (3.1)–(3.2) becomes:

$$Y = \begin{pmatrix} Z_\perp & W \end{pmatrix} \begin{pmatrix} \pi_{11,n} & \pi_{12,n} \\ \pi_{21,n} & \pi_{22,n} \end{pmatrix} + V, \qquad \pi_{11,n} = \pi_{12,n}\beta + \gamma_n,$$

where $\pi_{12,n}$ is the effect of the instruments (kindergarten classroom indicators) on the endogenous variable (kindergarten test score), and γ_n is the direct effect of the instruments on the outcome—the effect of first-grade teachers on first-grade test scores.

Under Normality, the model is still invariant to rotations of the instruments, with the maximal invariant still given by the statistics S and T. However, β can no longer be identified

from their expectations. In particular, we have:

$$\mathbb{E}[S] = \Omega, \qquad \mathbb{E}[T] = (k_n/n)\Omega + \Xi_n, \qquad \Xi_n = \Gamma\Lambda_n\Gamma',$$

where Γ is given in Equation (3.4), and

$$\Lambda_n = \begin{pmatrix} \gamma_n & \pi_{12,n} \end{pmatrix} (Z_\perp' Z_\perp / n) \begin{pmatrix} \gamma_n & \pi_{12,n} \end{pmatrix}.$$

This matrix governs the strength of instruments as well as the extent to which Assumption ER is violated. In particular, $\Lambda_{22,n} = \lambda_n / a'\Omega^{-1}a$ is proportional to Rothemberg's concentration parameter. $\Lambda_{11,n}$ measures the strength of the direct effects, and $\Lambda_{12,n}$ measures the strength of association between the first-stage coefficients and the direct effects. Assumption ER is equivalent to assuming $\Lambda_{11,n} = 0$, which implies that $\Lambda_{12,n} = 0$ also. These restrictions imply that $\Xi_n = \Lambda_{22,n} aa'$, which allows us to back out β from Ξ_n. However, if we don't assume that $\gamma_n = 0$, the only restriction on Λ_n is that it is positive semi-definite, and it is not possible to identify β. If we hope to identify β, it is therefore necessary to restrict the form of the direct effects in some way. Kolesár *et al.* (2011) show that β can be consistently estimated if the direct effects are orthogonal to the effects of the instruments on the endogenous variable in the following sense:

Assumption ODE (Orthogonal direct effects). $\Lambda_n \to \left(\begin{smallmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{smallmatrix}\right)$ for some $\Lambda_{11} \geq 0$ and $\Lambda_{22} > 0$.

In the context of the STAR example, this means that the effects of classroom assignment on test scores are orthogonal to the direct effects of classroom assignment on later outcomes, and it is satisfied by design if first-grade teachers are also randomly assigned. In other settings, this orthogonality requirement is still a substantive assumption that may or may not hold in practice, albeit weaker than the standard Assumption ER.

Kolesár *et al.* (2011) show that liml is not consistent under this assumption, but the bias-corrected two-stage least squares estimator (BTSLS, Donald and Newey, 2001) and the jackknife instrumental variables estimator (JIVE, Angrist *et al.*, 1999) are consistent,

provided $\alpha_\ell = 0$. If $\alpha_\ell > 0$, both estimators need some modification to maintain consistency. The original form of BTSLS is given by Equation (3.10), except with $m_{\min}$ replaced by $m_{\text{BTSLS}} = \frac{k_n-2}{n}\frac{n-k_n-\ell_n}{n-k_n+2}$:

$$\hat{\beta}_{\text{BTSLS}} = \frac{T_{12} - m_{\text{BTSLS}} S_{12}}{T_{22} - m_{\text{BTSLS}} S_{22}}.$$

Its many-exogenous-regressors-robust modification uses $m_{\text{MBTSLS}} = k_n/n$:

$$\hat{\beta}_{\text{MBTSLS}} = \frac{T_{12} - (k_n/n) S_{12}}{T_{22} - (k_n/n) S_{22}}.$$

Kolesár *et al.* (2011) also show that this estimator has smaller asymptotic variance than (modified) JIVE. This raises the question whether MBTSLS has the smallest asymptotic variance among estimators robust to the presence of direct effects.

3.5.2 Generalizing the RE framework to allow for direct effects

I will now generalize the random effects framework to answer this question and provide additional insight into the relationship between LIML and MBTSLS. In particular, I will model the normalized direct effects $\gamma^* = (Z_\perp' Z_\perp/n)^{1/2}\gamma_n$ as random and uncorrelated with the normalized first-stage coefficients $\pi^*_{12,n} = (Z_\perp' Z_\perp/n)^{1/2}\pi_{12,n}$:

$$\begin{pmatrix} \gamma_n^* \\ \pi_{12,n}^* \end{pmatrix} \sim N(0, \Lambda \otimes I_{k_n}/k_n), \qquad \Lambda = \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & \Lambda_{22} \end{pmatrix}. \tag{3.24}$$

The motivation for this prior is that if we view the coefficients $\pi^*_{12,n}$ and γ_n^* as random, then we can interpret the orthogonality assumption $\Lambda_{12,n} \to 0$ as saying that $\pi^*_{12,n}$, the random effects in the first stage, are uncorrelated with γ_n^*, the random effects in the structural equation. If $\Lambda_{11} = 0$, then the prior reduces to the random effects prior (3.16), since $\pi^*_{12,n} = \eta_n/\sqrt{a'\Omega^{-1}a}$.

Another possibility might be to try to modify the limited information likelihood so that it delivers an estimator under Assumptions MI and ODE. However, it is hard to incorporate the Assumption that $\Lambda_{12,n} \to 0$ into the likelihood, while allowing $\Lambda_{12,n}$ to differ from zero

in the sample.

The uncorrelated random effects (URE) likelihood based on integrating the density for the sufficient statistics $\hat{\Gamma}\Pi_1$ and S with respect to the URE prior (3.24) is given by (see Appendix C.3 for derivation)

$$\mathcal{L}_{\text{URE},n}(\beta, \Lambda_{11}, \Lambda_{22}, \Omega) = |\Omega|^{-(n-k_n-\ell_n)/2} e^{-\frac{n-k_n-\ell_n}{2}\operatorname{tr}(\Omega^{-1}S)} D(\beta, \Lambda_{11}, \Lambda_{22}, \Omega)^{-k_n/2} \cdot e^{-\frac{k_n}{2D(\beta,\Lambda_{11},\Lambda_{22},\Omega)}\left(\frac{k_n}{n}|\Omega|\operatorname{tr}(\Omega^{-1}T)+\Lambda_{22}b'Tb+\Lambda_{11}T_{22}\right)}, \quad (3.25)$$

where $D(\beta, \Lambda_{11}, \Lambda_{22}, \Omega)$ is the determinant of $\mathbb{E}[T] = \Omega + (k_n/n)\Gamma\Lambda\Gamma'$:

$$D(\beta, \Lambda_{11}, \Lambda_{22}, \Omega) = \frac{k_n}{n}\Lambda_{22}b'\Omega b + \frac{k_n}{n}\Lambda_{11}\Omega_{22} + \Lambda_{11}\Lambda_{22} + \frac{k_n^2}{n^2}|\Omega|.$$

The next proposition demonstrates that, as was the case in Proposition 3.2, the consistency properties of the maximum URE likelihood estimators do not rely on the URE prior (3.24):

Proposition 3.6. *Consider the model* (3.1)– (3.2).

(i) *Suppose that* $T_{22} \geq (k_n/n)S_{22}$. *Then the maximum likelihood estimator based on the* URE *likelihood* (3.25) *is given by:*

$$\hat{\beta}_{\text{URE}} = \frac{T_{12} - m_{\text{URE}}S_{12}}{T_{22} - m_{\text{URE}}S_{22}}, \qquad m_{\text{URE}} = \min\{m_{\min}, k_n/n\},$$

$$\hat{\Lambda}_{11,\text{URE}} = \max\left\{\hat{b}'_{\text{MBTSLS}}\left(T - \frac{k_n}{n}\hat{\Omega}_{\text{URE}}\right)\hat{b}_{\text{MBTSLS}}, 0\right\},$$

$$\hat{\Lambda}_{22,\text{URE}} = \begin{cases} T_{22} - \frac{k_n}{n}\hat{\Omega}_{22,\text{URE}} & \text{if } m_{\min} \geq k_n/n, \\ \frac{\hat{\lambda}_{\text{RE}}}{\hat{a}'_{\text{RE}}\hat{\Omega}_{\text{RE}}^{-1}\hat{a}_{\text{RE}}} & \text{otherwise.} \end{cases}$$

$$\hat{\Omega}_{\text{URE}} = \begin{cases} S & \text{if } m_{\min} \geq k_n/n, \\ \hat{\Omega}_{\text{RE}} & \text{otherwise.} \end{cases}$$

(ii) *Under Assumptions MI, N and ODE,* $(\hat{\beta}_{\text{URE}}, \hat{\Lambda}_{11,\text{URE}}, \hat{\Lambda}_{22,\text{URE}}, \hat{\Omega}_{\text{URE}}) \xrightarrow{p} (\beta, \Lambda_{11}, \Lambda_{22}, \Omega)$.

The key result in the proposition is that if $m_{\min} \geq k_n/n$, so that the URE estimate of Λ_{11} is positive, then $\hat{\beta}_{\text{URE}} = \hat{\beta}_{\text{MBTSLS}}$. Otherwise, if $m_{\min} < k_n/n$, then the likelihood is maximized at the boundary of the parameter space for Λ_{11}, and the URE estimates coincide with the

random effects estimates. If $\Lambda_{11} > 0$, then $m_{\min} \xrightarrow{p} \alpha_k + \min \operatorname{eig}(\Sigma^{-1}\Lambda) > \alpha_k$, so that $\hat{\beta}_{\text{URE}} = \hat{\beta}_{\text{MBTSLS}}$ in large samples. The motivation for introducing the MBTSLS estimator in Kolesár *et al.* (2011) was to modify the original Donald-Newey BTSLS estimators to make it consistent when $\alpha_\ell > 0$. Proposition 3.6 provides a maximum-likelihood motivation for this version of BTSLS.

To provide some insight into this result, consider a modification of the minimum distance estimator in Section 3.4.2 that allows for direct effects. Under the URE prior (3.24), we have:

$$\mathbb{E}[S] = \Omega, \qquad \mathbb{E}[T - k_n/nS] = \Xi, \qquad \Xi = \begin{pmatrix} \Lambda_{11} + \Lambda_{22}\beta^2 & \Lambda_{22}\beta \\ \Lambda_{22}\beta & \Lambda_{22} \end{pmatrix}.$$

Given some positive semi-definite weight matrix $\hat{W}_n$, the corresponding minimum distance objective function is given by:

$$\mathcal{Q}_{\text{simp},n}(\beta, \Lambda_{11}, \Lambda_{22}, \Omega; \hat{W}_n) = \operatorname{vech}(T - (k_n/n)S - \Xi)' \hat{W}_n \operatorname{vech}(T - (k_n/n)S - \Xi). \quad (3.26)$$

Under assumption ER, the reduced form coefficients are proportional to each other, $\pi_{11,n} = \pi_{12,n}\beta$. Consequently, $\Lambda_{11} = 0$, the matrix Ξ is reduced rank, and there are two sources of information for estimating β:

$$M_{11} = M_{12}\beta, \quad (3.27a)$$

$$M_{12} = M_{22}\beta. \quad (3.27b)$$

Weighting these two sources of identification using the weight matrix $\hat{W}_{\text{SIMP,RE}}$ given in Lemma 3.2 yields the minimum distance estimator $\hat{\beta}_{\text{LIML}}$, which is efficient under Normality. The price for its efficiency is that if $\Lambda_{11} > 0$, then Equation (3.27a) does not hold, which makes $\hat{\beta}_{\text{LIML}}$ sensitive to violations of the exclusion restriction. The same conclusion applies to the efficient minimum distance estimator $\hat{\beta}_{\text{EMD}}$.

On the other hand, MBTSLS does not restrict M_{11} in any way, and only uses (3.27b) to identify β. The model is exactly identified, and the weight matrix does not matter—the

minimum distance estimator will estimate β from (3.27b) by replacing the M_{11} and M_{12} by their consistent estimates. MBTSLS can therefore be viewed as a minimum distance estimator that puts no restrictions on the reduced form. Instead of assuming that the elements of $\pi^*_{12,n}$ and $\pi^*_{11,n}$ are proportional to each other, it only assumes that the proportionality holds "on average", in the sense of Equation (3.27b).

Finally, the URE estimator, like MBTSLS does not restrict Λ_{11} to be zero. It does, however, restrict it to be positive, viewing Λ as a covariance matrix of $(\gamma^*_j, \pi^*_{12,j})$. If the objective function (3.26) is minimized on the boundary of the parameter space of Λ_{11}, then it delivers the same estimates as the random effects estimator since it uses the same weight matrix. If the objective function is minimized in the interior, then the estimate of β will coincide with the MBTSLS estimator.

The next proposition summarizes these results:

Proposition 3.7. *Suppose that* $\operatorname{tr}(S^{-1}T) > 2k_n/n$. *Then the minimum distance estimator based on minimizing the objective function* (3.26) *with respect to the weight matrix* $\hat{W}_{\text{SIMP,RE}} = D_2'(S^{-1} \otimes S^{-1})D_2$ *is given by:*

(i) $(\hat{\beta}_{\text{LIML}}, 0, \hat{\Lambda}_{\text{RE},22})$ *if* Λ_{11} *is restricted to be zero;*

(ii) $(\hat{\beta}_{\text{URE}}, \hat{\Lambda}_{\text{URE},11}, \hat{\Lambda}_{\text{URE},22})$ *if* Λ_{11} *is restricted to be non-negative; and*

(iii) $(\hat{\beta}_{\text{MBTSLS}}, \hat{b}'_{\text{MBTSLS}}\left(T - (k_n/n)\hat{\Omega}_{\text{URE}}\right)\hat{b}_{\text{MBTSLS}}, T_{22} - (k_n/n)S_{22})$ *if* Λ_{11} *is unrestricted.*

The price for the extra robustness of MBTSLS and URE estimators of β is that they do not use the information contained in (3.27a) when the exclusion restriction holds, which results in larger asymptotic mean squared error than that of LIML and EMD. The next proposition uses the results Andrews (1999, 2002) on estimating parameters on the boundary to quantify the efficiency loss when Assumption ER does hold:

Proposition 3.8. *Consider the model* (3.1)–(3.2). *Under Assumptions ER, N, and MI:*

$$\sqrt{n}\left(\hat{\beta}_{\text{URE}} - \beta\right) \Rightarrow \sqrt{V_{\text{LIML},N}}\mathcal{Z}_2 + \frac{\sqrt{2\tau}\Sigma_{12}}{\Lambda_{22}}\max(\mathcal{Z}_1, 0), \qquad \mathcal{Z} \sim \mathcal{N}_2(0, I_2), \tag{3.28}$$

where $V_{\text{LIML},N} = \frac{\tau}{\Lambda_{22}^2}[\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2] + \Sigma_{11}/\Lambda_{22}$ *and* $\tau = \frac{\alpha_k(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell}$.

The asymptotic distribution is non-standard, and since $\mathbb{E}\max(\mathcal{Z}_1, 0) > 0$, the URE estimator is asymptotically biased. In comparison, recall from Equation (3.11) that the asymptotic variance of $\hat{\beta}_{\text{LIML}}$ under Assumptions ER, N, and MI is given by

$$\sqrt{n}\left(\hat{\beta}_{\text{LIML}} - \beta\right) \Rightarrow \sqrt{V_{\text{LIML},N}}\mathcal{Z}_2.$$

Therefore, the asymptotic efficiency loss of URE under Normality is captured by the second term in (3.28). If $\Sigma_{12} = 0$, then the efficiency loss is zero, and unless Λ_{22} is very small, the efficiency loss as measured by the mean squared error will be relatively small. Finally, Kolesár *et al.* (2011) show that for MBTSLS,

$$\sqrt{n}\left(\hat{\beta}_{\text{MBTSLS}} - \beta\right) \Rightarrow \sqrt{V_{\text{LIML}}}\mathcal{Z}_2 + \frac{\sqrt{2\tau}\Sigma_{12}}{\Lambda_{22}}\mathcal{Z}_1. \tag{3.29}$$

The difference between this expression and the asymptotic distribution on URE is that the term $\max(\mathcal{Z}_1, 0)$ in Equation 3.28 has been replaced by $\mathcal{Z}_1$. Lovell and Prescott (1970, Section 4) were the first ones to point out that this increases the asymptotic mean squared error.

If, on the other hand, Λ_{11} is bounded away from zero, then asymptotically the restriction that Λ_{11} has to be positive does not bind in large samples, and the asymptotic distribution of $\hat{\beta}_{\text{URE}}$ will coincide with that of $\hat{\beta}_{\text{MBTSLS}}$. To derive a specific distributional result however, Assumptions MI and ODE need to be strengthened. The problem is that even if $\Lambda_{12,n} \to 0$, the possibility that $\Lambda_{12,n} \neq 0$ affects the sampling distribution. Under assumption 3.24, when the URE likelihood is correctly specified, Kolesár *et al.* (2011) show that for the case when $\Lambda_{11} > 0$

$$\sqrt{n}\left(\hat{\beta}_{\text{URE}} - \beta\right) \Rightarrow \mathcal{N}\left(0, V_{\text{URE}}\right), \tag{3.30}$$

where

$$V_{\text{URE}} = \frac{1}{\Lambda_{22}^2}\left(\Lambda_{22}\Sigma_{11} + \frac{(1-\alpha_\ell)\alpha_k}{1-\alpha_k-\alpha_\ell}(\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2) + \Lambda_{11}\Omega_{22} + \Lambda_{11}\Lambda_{22}/\alpha_k\right). \tag{3.31}$$

There are two additional variance terms compared to the expression in Equation (3.29). If instead of Assumption 3.24, I assumed that $\Lambda_{12,n} = 0$, then the last term $\Lambda_{11}\Lambda_{22}/\alpha_k$ would

not appear in the variance expression. The last term also indicates that it is necessary that $\alpha_k > 0$ for the asymptotic variance to be finite. This is similar to clustering setups where the number of clusters needs to increase with the sample size. The construction of asymptotically valid confidence intervals might still be possible even with a small number of instruments by finding the asymptotic distribution of appropriately scaled t-statistics and inverting the associated t-tests (see Hansen (2007) and Donald and Lang (2007) for similar results in the clustering literature).

It is clear from the minimum distance representation (3.26) that if Λ_{11} is bounded away from zero, the model is exactly identified, so that trivially, the URE estimators of all parameters are efficient in the class of minimum distance estimators since the weight matrix $\hat{W}_n$ does not matter. Moreover, since the URE model is a member of the exponential family, it is easy to show by standard Taylor expansion arguments (see, for example van der Vaart, 1998, Chapter 7) that the uncorrelated random effects model is locally asymptotically Normal under Assumptions MI and 3.24. Therefore, if Λ_{11} is bounded away from zero, so that the local parameter space is unrestricted, then $(\hat{\beta}_{\text{URE}}, \hat{\Lambda}_{11,\text{URE}}, \hat{\Lambda}_{22,\text{URE}}, \hat{\Omega}_{\text{URE}})$ is regular, it coincides with the unrestricted estimator given by Proposition 3.7 (iii), and it is asymptotically efficient among regular estimators. If $\Lambda_{11} = 0$, so that the local parameter space for Λ_{11} comprises the positive part of the real line and is unrestricted for the remaining parameters, then, asymptotically, $(\hat{\beta}_{\text{URE}}, \hat{\Lambda}_{11,\text{URE}}, \hat{\Lambda}_{22,\text{URE}}, \hat{\Omega}_{\text{URE}})$ has the same properties as the maximum likelihood estimator of Normal means with known variance, with one of the means (corresponding to the local parameter space for Λ_{11}) restricted to be non-negative.

These results make $\hat{\beta}_{\text{URE}}$ an attractive robust choice of estimator. Unlike LIML, URE is robust to $\Lambda_{11} > 0$, in which case it is efficient and coincides with MBTSLS in large sample. When no direct effects are present, its asymptotic mean-square error is slightly higher than that of LIML, but lower than that of MBTSLS.

One factor complicating inference about β using $\hat{\beta}_{\text{URE}}$ that is valid uniformly over the parameter space for Λ_{11} is the non-standard form of its asymptotic distribution when $\Lambda_{11} = 0$. There are several possible approaches that address this issue. I discuss two of

them (see Andrews, 1999, for discussion of a version of bootstrap and subsampling).

The first approach is based on the observation that the conventional asymptotic standard errors based on the assumption that no parameters are on the boundary (i.e. standard errors based on Equation (3.31)) yield conservative confidence intervals when, in fact $\Lambda_{11} = 0$ (Andrews, 1999, p. 1369). Under assumption 3.24, the model for $\hat{\Pi}_1^*$ and S constitutes an exponential family, and the two natural ways of estimating V_{URE} in Equation (3.31)—using the (1,1) element of the inverse Hessian, evaluated at the maximum likelihood estimates, and using the inverse of the information matrix evaluated at the maximum likelihood estimates—coincide. In both cases, the estimator of the asymptotic variance is given by (3.31) with the parameters Σ, Λ_{22} and Λ_{11} replaced by their maximum likelihood estimates, with the maximum likelihood estimate of Σ given by:

$$\hat{\Sigma}_{\text{RE}} = \hat{\Gamma}_{\text{RE}}^{-1} \hat{\Omega}_{\text{RE}} \hat{\Gamma}_{\text{RE}}^{-1}, \qquad \hat{\Gamma}_{\text{RE}} = \begin{pmatrix} e_1 & \hat{a}_{\text{RE}}. \end{pmatrix}$$

This estimator will be consistent for V_{URE} under Assumptions MI and ODE.

The second approach suggested by Andrews (1999) is to do a pre-test of the hypothesis $H_0 \colon \Lambda_{11} = 0$ against $H_1 \colon \Lambda_{11} > 0$ to determine if the true parameter Λ_{11} is at the boundary with critical values chosen such that the pre-test is consistent as $n \to \infty$. If the test rejects, then we conclude that Λ_{11} is not at the boundary and we use Hessian-based standard errors. Otherwise, we assume that $\Lambda_{11} = 0$. In this case, we use the asymptotic distribution (3.28) to obtain confidence intervals. In particular, use plug-in estimators of V_{LIML} and $\sigma_1 = \sqrt{2\tau}\Sigma_{12}/\Lambda_{22}$ based on the URE estimates. Although the quantiles of $\hat{V}_{\text{LIML}} \mathcal{Z}_2 + \hat{\sigma}_1 \mathcal{Z}_1$ cannot be obtained in closed form, they can easily be simulated by taking draws of $(\mathcal{Z}_1, \mathcal{Z}_2)$. The pre-test used in this approach is, in fact, equivalent to some consistent test of overidentifying restrictions. I discuss these tests in detail in the next Section, in which I show that one possibility is to reject whenever $m_{\min} - k_n/n$ is greater than some fixed constant.

3.6 Tests of overidentifying restrictions

Assumption ER imposes a proportionality restriction on the reduced form (3.3) that $\pi_{11,n} = \pi_{12,n}\beta$. If Assumption ER does not hold, the reduced-form coefficients are unrestricted. A variety of tests of this restriction that work under the standard asymptotics that hold k_n and ℓ_n fixed have been proposed in the literature. First, I will discuss the robustness of three such tests to the presence of many instruments and many exogenous regressors. I will then relate these tests to tests based on the random effects likelihood and the minimum distance objective functions.

The most popular test, due to Sargan (1958), is based on the observation that the nR^2 from regressing the estimated residuals in the structural equation (3.1a) on the instruments and exogenous variables is asymptotically distributed according to χ^2_{k-1} under Assumption ER and standard asymptotics that hold the number of instruments and exogenous regressors fixed, so that $k_n = k, \ell_n = \ell$. If LIML is used to estimate β and δ_n, the estimated residuals can be written as $Y_\perp \hat{\beta}_{\text{LIML}}$, and consequently, the R^2 is given by

$$\hat{J}_{\text{S}} = \frac{\hat{b}'_{\text{LIML}} T \hat{b}_{\text{LIML}}}{\hat{b}'_{\text{LIML}} (Y'_\perp Y_\perp / n) \hat{b}_{\text{LIML}}} = \frac{m_{\min}}{1 - k_n/n - \ell_n/n + m_{\min}}.$$

The Sargan test therefore rejects if $n\hat{J}_{\text{S}}$ is greater than $q^{\chi^2_k}_{\text{ns}}$, the $1 - \text{ns}$ quantile of a χ^2_{k-1} distribution where ns denotes the desired nominal size.

A closely related alternative is the generalized likelihood ratio test based on the limited information likelihood of Anderson and Rubin (1949). The test statistic is given by $n\hat{J}_{\text{AR}}$, where $\hat{J}_{\text{AR}} = n \log(n m_{\min} / (n - k_n - \ell_n) + 1)$. It is also asymptotically distributed according to χ^2_{k-1} under the null and standard asymptotics.

Third, Cragg and Donald (1993) suggest a test based on the minimum distance objective function (3.21). They show that the minimum of the objective function is given by

$$\hat{J}_{\text{CD}} = \frac{\hat{b}'_{\text{LIML}} T \hat{b}_{\text{LIML}}}{\hat{b}'_{\text{LIML}} S \hat{b}_{\text{LIML}}} = m_{\min}.$$

Compared to the Sargan test statistic, $\hat{J}_{\text{CD}}$ replaces $(Y'_\perp Y_\perp)/n$ by S in the denominator. Cragg

and Donald (1993) also show that $nm_{\min} \Rightarrow \chi^2_{k-1}$ under standard asymptotics.

All three tests are equivalent in the sense that they all reject for large values of $m_{\min}$. Therefore, the only difference between them in finite samples is how well the chi-squared approximation controls size in each case. While under standard asymptotics their asymptotic distributions coincide and therefore do not provide any guidance as to which test has the best size control, allowing for $\alpha_k, \alpha_\ell > 0$ reverses this conclusion:

Lemma 3.3. *Consider the model* (3.1)–(3.2)*. Then, under Assumptions ER and MI:*

$$n^{1/2}\left(\hat{J}_{\text{S}} - \tfrac{\alpha_k}{1-\alpha_\ell}\right) \Rightarrow \mathcal{N}\left(0, \tfrac{2\alpha_k(1-\alpha_k-\alpha_\ell)}{(1-\alpha_\ell)^3}\right), \tag{3.32}$$

$$n^{1/2}\left(\hat{J}_{\text{AR}} - \log\left(\tfrac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}\right)\right) \Rightarrow \mathcal{N}\left(0, 2\tau/(1-\alpha_\ell)^2\right), \tag{3.33}$$

$$n^{1/2}\left(\hat{J}_{\text{CD}} - \alpha_k\right) \Rightarrow \mathcal{N}(0, 2\tau), \tag{3.34}$$

where $\tau = \frac{\alpha_k(1-\alpha_\ell)}{1-\alpha_k-\alpha_\ell}$*. Moreover, if* $\alpha_k > 0$*,*

$$\mathbb{P}\left(n\hat{J}_{\text{S}} \geq q^{\chi^2_{k_n-1}}_{ns}\right) \to \begin{cases} \Phi(\Phi^{-1}(ns)/\sqrt{1-\alpha_k}) & \text{if } \alpha_\ell = 0, \\ 1 & \text{otherwise.} \end{cases}$$

$$\mathbb{P}\left(n\hat{J}_{\text{AR}} \geq q^{\chi^2_{k_n-1}}_{ns}\right) \to 1,$$

$$\mathbb{P}\left(n\hat{J}_{\text{CD}} \geq q^{\chi^2_{k_n-1}}_{ns}\right) \to \Phi\left(\Phi^{-1}(ns)\sqrt{(1-\alpha_k-\alpha_\ell)/(1-\alpha_\ell)}\right),$$

where $\Phi(\cdot)$ *is the cdf of a standard Normal distribution.*

Anatolyev and Gospodinov (2011) and Anatolyev (2011) derive the results for the Sargan test. The results for the Anderson-Rubin overidentification test and the Cragg-Donald test are new.

When $\alpha_k > 0$ and $\alpha_\ell = 0$, the Sargan test is mildly conservative. With $\alpha_k = 0.1$ for example, the asymptotic size of the test with nominal size 0.05 is given by 0.04. Anatolyev and Gospodinov (2011) therefore propose an adjustment to the critical value of the Sargan test to match the asymptotic size with the nominal size—instead of using the $q^{\chi^2_k}_{\text{s}}$ critical value, they suggest using $q^{\chi^2_{k_n-1}}_{\Phi(\sqrt{1-\alpha_k}\Phi^{-1}(\text{s}))}$, which will have the correct asymptotic size. As the Lemma demonstrates, the problem with this solution is that it breaks down when

$\alpha_\ell > 0$. Furthermore, it is no longer possible to adjust the critical value to correct the asymptotic size because the test statistic is centered at the wrong value—$\alpha_k/(1-\alpha_\ell)$ rather than $\alpha_k = \mathbb{E}[\chi^2_{k_n}/n]$. Similar conclusions apply to the Anderson-Rubin overidentification test.

The Cragg-Donald test is also size-distorted, although the distortion is rather small. With $\alpha_k = \alpha_\ell = 0.1$ for example, the asymptotic size of the test with nominal size 0.05 is given by 0.07. Moreover, we can apply the Anatolyev and Gospodinov (2011) adjustment to the critical value to correct the size distortion. In particular, comparing $nm_{\min}$ against the $\Phi(\sqrt{(1-\alpha_\ell)/(1-\alpha_k-\alpha_\ell)}\Phi^{-1}(\text{ns}))$ quantile of the $\chi^2_{k_n-1}$ distribution will yield a critical value that will control size under standard as well as many-instrument asymptotics.

An alternative to size-correcting existing tests of overidentification to make them robust to the presence of many instruments is to make use of the random-effects framework directly. In the URE model, absence of direct effects of the instruments is equivalent to $\Lambda_{11} = 0$, in which case the model reduces to the random-effects model. If the exclusion restriction fails, then $\Lambda_{11} > 0$, and the matrix M is no longer reduced rank. In this case, the assumption that $\Lambda_{12} = 0$ is not restrictive as it doesn't restrict the distribution of the reduced-form coefficients—it only serves to identify β in the URE model (another consequence of this fact is that $\Lambda_{12} = 0$ is not a testable assumption). Therefore, testing for overidentifying restrictions in the URE model is equivalent to testing $H_0\colon \Lambda_{11} = 0$ against $H_1\colon \Lambda_{11} > 0$. The next lemma gives the form of two such tests. The first test is the likelihood ratio test based on the URE likelihood. The second test is a J-test based on the minimum distance objective function (3.19).

Lemma 3.4. *The generalized likelihood ratio test statistic for overidentifying restrictions based on the URE likelihood is given by*

$$\hat{J}_{\text{RE}} = \begin{cases} 0 & \text{if } m_{\min} \leq k_n/n, \\ (n-\ell_n)\log\left(\frac{n-k_n-\ell_n+nm_{\min}}{(n-\ell_n)}\right) - k_n\log\left(\frac{nm_{\min}}{k_n}\right) & \text{otherwise.} \end{cases}$$

The J-test statistic of overidentifying restrictions based on the minimum distance objective function

(3.19) *is given by*

$$\hat{J}_{\text{MD}} = \begin{cases} 0 & \text{if } m_{\min} \leq k_n/n, \\ \frac{1-k_n/n-\ell_n/n}{k_n/n(1-\ell_n/n)} \left(m_{\min} - k_n/n\right)^2 & \text{otherwise.} \end{cases}$$

Again, the tests are equivalent to the Sargan, Anderson-Rubin and Cragg-Donald tests of overidentification in the sense that all tests reject for large values of $m_{\min}$. Moreover, the $\hat{J}_{\text{MD}}$ test is equivalent to the one-sided Cragg-Donald test based on the approximation (3.34), so that the two minimum distance objective functions (3.19) and (3.21) deliver the same test statistic. These results suggest that the preferred test for overidentifying restrictions is given by the size-corrected Cragg-Donald test.

3.7 Conclusion

In this paper, I outlined an integrated likelihood approach to inference in the instrumental variables model when the number of instruments is large. This approach addresses the incidental parameter problem that the large number of instruments create. It it principled and unified, as it explicitly uses an invariance argument to deal with the incidental parameters and it is based on a well-motivated and well-behaved objective function. I show that this integrated likelihood coincides with the random effects likelihood of Chamberlain and Imbens (2004), and that the maximum likelihood estimator of β coincides with LIML. Moreover, maximizing this integrated likelihood is equivalent to minimizing a minimum distance objective function. I use this equivalence to show that when the reduced-form errors are not Normal, a minimum distance estimator with respect to an efficient weight matrix is more efficient than LIML. Finally, I generalize the random effects likelihood to allow instruments to have direct effects on the outcome that are orthogonal to the effects of the instruments on the endogenous variable. I show that the resulting maximum likelihood estimator, the uncorrelated random effects estimator, is a mixture between LIML and the bias-corrected two stage least squares estimator. It shares the robustness of the bias-corrected

two stage least squares (BTSLS) estimator to violations of the exclusion restriction, while, like LIML, it exploits the exclusion restriction when it holds, so that it is more efficient than BTSLS.

References

Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, **113** (2), 231–263.

Abramowitz, M. and Stegun, I. A. (eds.) (1965). *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover.

Ackerberg, D. A. and Devereux, P. J. (2009). Improved Jive estimators for overidentified linear models with and without heteroskedasticity. *Review of Economics and Statistics*, **91** (2), 351–362.

Aizer, A. and Doyle, Jr., J. J. (2011). Effects of Juvenile Incarceration: Evidence from Randomly-Assigned Judges.

Alonso-Borrego, C. and Arellano, M. (1999). Symmetrically normalized instrumental-variable estimation using panel data. *Journal of Business & Economic Statistics*, **17** (1), 36–49.

Anatolyev, S. (2011). Instrumental variables estimation and inference in the presence of many exogenous regressors.

— and Gospodinov, N. (2011). Specification testing in models with many instruments. *Econometric Theory*, **27** (2), 427–441.

Anderson, T. W., Kunitomo, N. and Matsushita, Y. (2010). On the asymptotic optimality of the LIML estimator with possibly many instruments. *Journal of Econometrics*, **157** (2), 191–204.

— and Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, **20** (1), 46–63.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, **67** (6), 1341–1383.

— (2002). Generalized Method of Moments Estimation When a Parameter Is on a Boundary. *Journal of Business & Economic Statistics*, **20** (4), 530–544.

—, Moreira, M. J. and Stock, J. H. (2006). Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression. *Econometrica*, **74** (3), 715–752.

—, — and — (2008). Efficient two-sided nonsimilar invariant tests in IV regression with weak instruments. *Journal of Econometrics*, **146** (2), 241–254.

Angrist, J. D. and Imbens, G. W. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity. *Journal of the American Statistical Association*, **90** (430), 431–442.

—, — and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, **14** (1), 57–67.

—, — and Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, **91** (434), 444–455.

— and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106** (4), 979–1014.

— and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press.

Ashley, R. (2009). Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis. *Journal of Applied Econometrics*, **24** (2), 325–337.

Basmann, R. L. (1957). A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, **25** (1), 77–83.

Bekker, P. A. (1994). Alternative Approximations to the Distributions of Instrumental Variable Estimators. *Econometrica*, **62** (3), 657–681.

— and Crudu, F. (2012). Symmetric Jackknife Instrumental Variable Estimation.

— and van der Ploeg, J. (2005). Instrumental variable estimation based on grouped data. *Statistica Neerlandica*, **59** (3), 239–267.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. B. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, (forthcoming).

Berkowitz, D., Caner, M. and Fang, Y. (2008). Are "Nearly Exogenous Instruments" reliable? *Economics Letters*, **101** (1), 20–23.

Bound, J., Jaeger, D. A. and Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, **90** (430), 443 – 450.

Caner, M. (2007). Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Many Moment Asymptotics.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, (in press).

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, **18** (1), 5–46.

— (2007). Decision theory applied to an instrumental variables model. *Econometrica*, **75** (3), 609–652.

— and Imbens, G. W. (2004). Random Effects Estimators with Many Instrumental Variables. *Econometrica*, **72** (1), 295–306.

— and Moreira, M. J. (2009). Decision Theory Applied to a Linear Panel Data Model. *Econometrica*, **77** (1), 107–133.

Chao, J. C., Hausman, J. A., Newey, W. K., Swanson, N. R. and Woutersen, T. (2010). Testing Overidentifying Restrictions with Many Instruments and Heteroscedasticity.

— and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, **73** (5), 1673–1692.

—, —, Hausman, J. A., Newey, W. K. and Woutersen, T. (2012). Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments. *Econometric Theory*, **12** (1), 42–86.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W. and Yagan, D. (2011). How does your kindergarten classroom affect your earnings? *Quarterly Journal of Economics*, **126** (4), 1593–1660.

Chioda, L. and Jansson, M. (2009). Optimal Invariant Inference When the Number of Instruments Is Large. *Econometric Theory*, **25** (3), 793–805.

Conley, T., Hansen, C. and Rossi, P. (2012). Plausibly exogenous. *The Review of Economics and Statistics*, **94** (1), 260–272.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, **49** (1), 1–39.

Cragg, J. G. and Donald, S. G. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, **9** (2), 222–240.

Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.

Dobbie, W. and Fryer, R. G. (2011). Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children's Zone. *American Economic Journal: Applied Economics*, **3** (3), 158–187.

— and Song, J. (2012). Debt Relief and Debtor Outcomes: Measuring the Effects of Consumer Bankruptcy Protection.

Donald, S. G. and Lang, K. (2007). Inference with Difference-in-Differences and Other Panel Data. *The Review of Economics and Statistics*, **89** (2), 221–233.

— and Newey, W. K. (2001). Choosing the Number of Instruments. *Econometrica*, **69** (5), 1161–1191.

Eaton, M. L. (1989). *Group invariance applications in statistics*, *Regional conference series in Probability and Statistics*, vol. 1. Hayward, California: Institute of Mathematical Statistics.

Fisher, F. M. (1961). On the cost of approximate specification in simultaneous equation estimation. *Econometrica*, **29** (2), 139–170.

— (1966). The relative sensitivity to specification error of different k-class estimators. *Journal of the American Statistical Association*, **61** (314), 345–356.

— (1967). Approximate Specification and the Choice of a k-Class Estimator. *Journal of the American Statistical Association*, **62** (320), 1265–1276.

Flores, C. A. and Flores-Lagunes, A. (2010). Partial Identification of Local Average Treatment Effects with an Invalid Instrument.

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, **139** (1), 35–75.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica*, **45** (4), 939–953.

Gautier, E. and Tsybakov, A. B. (2011). High-dimensional instrumental variables regression and confidence sets.

Goldberger, A. S. and Olkin, I. (1971). A minimum-distance interpretation of limited-information estimation. *Econometrica*, **39** (3), 635–639.

Guggenberger, P. (2012). On the Asymptotic Size Distortion of Tests When Instruments Locally Violate the Exogeneity Assumption. *Econometric Theory*, **28** (2), 387–421.

Hahn, J. (2002). Optimal inference with many instruments. *Econometric Theory*, **18** (1), 140–168.

— and Hausman, J. A. (2005). IV Estimation with Valid and Invalid Instruments. *Annales d'Économie et de Statistique*, (79/80), 25–57.

Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics*, **141** (2), 597–620.

—, Hausman, J. A. and Newey, W. K. (2008). Estimation With Many Instrumental Variables. *Journal of Business and Economic Statistics*, **26** (4), 398–422.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C. and Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, **3** (2), 211–255.

Heckman, J. J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, **4** (5), 475–492.

— (1997). Instrumental variables: A study of implicit behavioral assumptions used in making program evaluations. *Journal of Human Resources*, **32** (3), 441–452.

—, Urzua, S. and Vytlacil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, **88** (3), 389–432.

— and Vytlacil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences of the United States of America*, **96** (8), 4730–4734.

— and — (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica*, **73** (3), 669–738.

Hillier, G. H. (1990). On the normalization of structural equations: Properties of direction estimators. *Econometrica*, **58** (5), 1181–1194.

Hirano, K., Imbens, G. W., Rubin, D. B. and Zhou, X.-H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, **1** (1), 69–88.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** (2), 467–475.

Keller, W. J. (1975). A new class of limited-information estimators for simultaneous equations systems. *Journal of Econometrics*, **3** (1), 71–92.

Kolesár, M. (2012). Random-Effects Approach to Inference With Many Instruments.

Kolesár, M., Chetty, R., Friedman, J. N., Glaeser, E. and Imbens, G. W. (2011). Identification and Inference with Many Invalid Instruments.

Kraay, A. (2008). Instrumental Variables Regressions with Honestly Uncertain Exclusion Restrictions.

Kunitomo, N. (1980). Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations. *Journal of the American Statistical Association*, **75** (371), 693–700.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, **95** (2), 391–413.

— (2002). Orthogonal Parameters and Panel Data. *Review of Economic Studies*, **69** (3), 647–666.

Lovell, M. C. and Prescott, E. (1970). Multiple regression with inequality constraints: Pretesting bias, hypothesis testing and efficiency. *Journal of the American Statistical Association*, **65** (330), 913–925.

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: Some properties and applications. *The Annals of Statistics*, **7** (2), 381–394.

— and — (1980). The elimination matrix: Some lemmas and applications. *SIAM Journal on Algebraic and Discrete Methods*, **1** (4), 422–449.

Moreira, M. J. (2003). A Conditional Likelihood Ratio Test for Structural Models. *Econometrica*, **71** (4), 1027–1048.

— (2009). A maximum likelihood method for the incidental parameter problem. *The Annals of Statistics*, **37** (6A), 3660–3696.

Morimune, K. (1983). Approximate distributions of k-class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica*, **51** (3), 821–841.

Nagar, A. L. (1959). The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations. *Econometrica*, **27** (4), 575–595.

Nagin, D. and Snodgrass, M. G. (2011). The Effect of Incarceration on Offending: Evidence from a Natural Experiment in Pennsylvania.

Nevo, A. and Rosen, A. M. (2012). Identification with Imperfect Instruments. *Review of Economics and Statistics*, **94** (3), 659–671.

Newey, W. K. and McFadden, D. L. (1994). Large sample estimation and hypothesis testing. In R. F. Engle and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 4, *Chapter 36*, Elsevier, pp. 2111–2245.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16** (1), 1–32.

Peiser, A. M. (1943). Asymptotic formulas for significance levels of certain distributions. *The Annals of Mathematical Statistics*, **14** (1), 56–62.

Phillips, G. D. A. and Hale, C. (1977). The Bias of Instrumental Variable Estimators of Simultaneous Equation Systems. *International Economic Review*, **18** (1), 219–228.

Phillips, P. C. B. (1983). Exact small sample theory in the simultaneous equations model. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of Econometrics*, Elsevier, vol. 1, pp. 449–516.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics, New York: Springer-Verlag.

Reinhold, S. and Woutersen, T. (2011). Endogeneity and Imperfect Instruments in Applied Work: Deriving Bounds in a Semiparametric Model.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. In Z. Griliches and M. D. Intriligator (eds.), *Handbook of econometrics*, vol. 2, *Chapter 15*, Elsevier, pp. 881–935.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, **26** (3), 393–415.

Sims, C. A. (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of Econometrics*, **95** (2), 443–462.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, **65** (3), 557–586.

Theil, H. (1961). *Economic Forecasts and Policy*. Amsterdam: Horth-Holland, 2nd edn.

— (1971). *Principles of Econometrics*. New York: John Wiley & Sons.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

van Hasselt, M. (2010). Many Instruments Asymptotic Approximations Under Nonnormal Error Distributions. *Econometric Theory*, **26** (02), 633–645.

Vytlacil, E. J. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica*, **70** (1), 331–341.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Yau, L. H. and Little, R. J. A. (2001). the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. *Journal of the American Statistical Association*, **96** (456), 1232–1244.

Zellner, A. (1970). Estimation of regression relationships containing unobservable independent variables. *International Economic Review*, **11** (3), 441–454.

Appendix A

Appendix to Chapter 1

First I define some notation and collect some basic results that I use throughout Appendices A.1 and A.2. I use the notation $\|\cdot\|_F$ to denote the Frobenius norm, so that for any matrix A, $\|A\|_F = \sqrt{\text{tr}(AA')}$. By the Cauchy-Schwarz inequality, we then have $\text{tr}(A'B) \leq \|A\|_F \|B\|_F$. The Frobenius norm is sub-multiplicative, so that $\|AB\|_F \leq \|A\|_F \|B\|_F$ for any matrices A and B. Let $\mathcal{Z}_n = \{Q_i, X_i\}_{i=1}^n$ denote the collection of covariates and instruments. Also, let

$$\mathbf{G}_n = (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n, \mathbf{W}_n)})^{-1}(\mathbf{H}_{(\mathbf{Z}_n, \mathbf{W}_n)} - \mathbf{D}_{(\mathbf{Z}_n, \mathbf{W}_n)}) - (\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n}). \tag{A.1}$$

Then $\hat{\mathbf{P}}_{\text{UJIVE}} = \mathbf{G}_n \mathbf{T}_n$.

A.1 Auxiliary Lemmata

Lemma A.1. *Let* $\tilde{P}_i^L = \mathbb{E}^*[T_i \mid Z_i, W_i] - \mathbb{E}^*[T_i \mid W_i]$, *and let* $\tilde{R}_i^L = \mathbb{E}^*[Y_i \mid Z_i, W_i] - \mathbb{E}^*[Y_i \mid W_i]$. *Then:*

(i) $\Xi_{12} = \mathbb{E}[Y_i \tilde{P}_i^L]$;

(ii) $\Xi_{12} = \mathbb{E}[T_i \tilde{R}_i^L]$;

(iii) $\Xi_{11} = \mathbb{E}[Y_i \tilde{R}_i^L]$; *and*

(iv) $\Xi_{22} = \mathbb{E}[T_i \tilde{P}_i^L]$.

Proof. Consider Part (i). Observe that

$$\begin{aligned}\mathbb{E}[Y_i \tilde{P}_i^L] &= \mathbb{E}[(Z_i'\pi_1 + W_i'\psi_1)'\tilde{P}_i^L] \\ &= \mathbb{E}[(Z_i'\pi_1 + W_i'\psi_1)'\tilde{Z}_i'\pi_2] \\ &= \mathbb{E}[\pi_1 Z_i \tilde{Z}_i \pi_2] = \mathbb{E}[\pi_1 \tilde{Z}_i \tilde{Z}_i \pi_2] \\ &= \Xi_{12},\end{aligned}$$

where the first line follows from Equation (1.3) and the fact that $\tilde{P}_i^L$ is linear in Z_i and W_i, the second line follows from $\tilde{P}_i^L = \tilde{Z}_i'\pi_1$, the third line follows from $\mathbb{E}[W_i\tilde{Z}_i] = 0$, and the last line follows by definition of Ξ_{12}. Parts (ii)–(iv) follow by similar arguments, using the substitutions $\tilde{R}_i^L = \tilde{Z}_i'\pi_1$ and $\tilde{P}_i^L = \tilde{Z}_i'\pi_2$. □

Lemma A.2. *Let $A_i = a(Q_i, X_i)$ be some function of the instruments and covariates such that $\mathbb{E}[A_i \mid X_i] = 0$. Then, under Assumptions IV and M,*

$$\mathbb{E}[Y_i A_i] = \int \sum_{j=1}^{J_x - 1} \alpha(p_{j,x}; x)(p_{j,x} - p_{j-1,x})\mathbb{E}[A_i \mid X_i = x, P_i > p_{j,x}]\mathbb{P}(P_i > p_{j,x} \mid X_i = x)\, \mathrm{d}F^X(x),$$

$$\mathbb{E}[T_i A_i] = \int \sum_{j=1}^{J_x - 1} (p_{j,x} - p_{j-1,x})\mathbb{E}[A_i \mid X_i = x, P_i > p_{j,x}]\mathbb{P}(P_i > p_{j,x} \mid X_i = x)\, \mathrm{d}F^X(x).$$

Proof. First consider $\mathbb{E}[Y_i A_i]$. By the Law of iterated expectations,

$$\begin{aligned}\mathbb{E}[Y_i A_i] &= \int \sum_j \sum_a a\, \mathbb{E}[Y_i \mid X_i = x, A_i = a, P_i = p_{j,x}]\mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x)\, \mathrm{d}F^X(x) \\ &= \int \sum_j \sum_a a\, \mathbb{E}[Y_i \mid X_i = x, P_i = p_{j,x}]\mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x)\, \mathrm{d}F^X(x)\end{aligned} \tag{A.2}$$

where the second line follows from the fact that under Assumptions IV and M, the conditional expectation of Y_i depends only on P_i and X_i. Using the substitution

$$\mathbb{E}[Y_i \mid X_i = x, P_i = p_{j,x}] = \mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] + \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}),$$

we can expand the expression (A.2) as

$$\begin{aligned}\mathbb{E}[Y_i A_i] &= \int \sum_j \sum_a a \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x})\mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x)\, \mathrm{d}F^X(x) + \\ &\quad + \int \mathbb{E}[A_i \mid X_i = x]\mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}]\, \mathrm{d}F^X(x) \\ &= \int \sum_j \sum_a a \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x})\mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x)\, \mathrm{d}F^X(x) \\ &= \int \sum_{j'=1}^{J_x - 1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{j > j'} \sum_a a \mathbb{P}(P_i = p_{j,x}, A_i = a \mid X_i = x)\, \mathrm{d}F^X(x),\end{aligned}$$

where the second line follows since $\mathbb{E}[A_i \mid X_i] = 0$ by assumption, and the last line follows from changing the

order of summation. Therefore, by definition of conditional expectation:

$$\begin{aligned}\mathbb{E}[Y_i A_i] &= \int \sum_{j=1}^{J_x-1} \alpha(p_{j,x}; x)(p_{j,x} - p_{j-1,x}) \sum_{j>j'} \mathbb{E}[A_i \mid P_i = p_{j,x}, X_{i=x}] \mathbb{P}(P_i = p_{j,x} \mid X_i = x) \, \mathrm{d}F^X(x) \\ &= \int \sum_{j=1}^{J_x-1} \alpha(p_{j,x}; x)(p_{j,x} - p_{j-1,x}) \mathbb{E}[A_i \mid X_i = x, P_i > p_{j,x}] \mathbb{P}(P_i > p_{j,x} \mid X_i = x) \, \mathrm{d}F^X(x).\end{aligned} \tag{A.3}$$

The expression for $\mathbb{E}[T_i A_i]$ can be derived using the same arguments, except that we substitute

$$\mathbb{E}[T_i \mid X_i = x, P_i = p_{j,x}] = p_{1,x} + \sum_{j'=1}^{j-1} (p_{j'+1,x} - p_{j',x}). \qquad \square$$

I use the following results from Chao *et al.* (2012) and Politis, Romano and Wolf (1999) to prove Lemma A.5 and Lemma A.6 below:

Lemma A.3 (Chao *et al.*, 2012, Lemma A.1). *Suppose that, conditional on some set of random variables $\mathcal{F}$, $\{(A_i, B_i)\}_{i=1}^n$ is independent a.s., where A_i and B_i are some scalars random variables. Let $\mathbf{H}$ be a symmetric idempotent matrix with rank K. Let $\mathbb{E}[A_i \mid \mathcal{F}] = \bar{a}_i$, $\mathbb{E}[B_i \mid \mathcal{F}] = \bar{b}_i$, and $\sigma_A^2 = \max_{i \leq n} \mathrm{var}(A_i \mid \mathcal{F})$, $\sigma_B^2 = \max_{i \leq n} \mathrm{var}(B_i \mid \mathcal{F})$. Then there exists a positive constant C such that*

$$\mathbb{E}\left[\left(\sum_i \sum_{j \neq i} (A_i H_{ij} B_j - \bar{a}_i H_{ij} \bar{b}_j) \right)^2 \,\middle|\, \mathcal{F} \right] \leq C(K\sigma_A^2 \sigma_B^2 + \sigma_A^2 \bar{b}'\bar{b} + \sigma_B^2 \bar{a}'\bar{a}).$$

Lemma A.4 (Lemma 1.3.2., Politis *et al.*, 1999). *Suppose $(A_{n,1}, \ldots, A_{n,n})$ is a triangular array of i.i.d. random variables, the nth row having distribution F_n^A. Assume F_n^A converges in distribution to F^A, and $\mathbb{E}[|A_{n,1}|] \to \mathbb{E}[|A|] < \infty$, as $n \to \infty$, where A is distributed according to F^A. Then $n^{-1} \sum_{i=1}^n A_{n,i} \to E[A]$ as $n \to \infty$.*

Lemma A.5. *Suppose that Assumptions R and MI hold. Then*

(i) $\mathbf{T}_n' \mathbf{G}_n \mathbf{T}_n / n = \mathbf{P}_n' \mathbf{G}_n \mathbf{P}_n / n + o_p(1)$*; and*

(ii) $\mathbf{Y}_n' \mathbf{G}_n \mathbf{T}_n / n = \mathbf{R}_n' \mathbf{G}_n \mathbf{P}_n / n + o_p(1)$,

where $\mathbf{G}_n$ is defined in Equation (A.1)*.*

Proof. I will prove Part (i), Part (ii) follows by similar arguments. Let $A_{n,i}^W = (1 - (\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1} T_{n,i}$, and let $A_{n,i}^{(Z,W)} = (1 - (\mathbf{H}_{(\mathbf{Z}_n, \mathbf{W}_n)})_{ii})^{-1} T_{n,i}$. Denote by $\bar{a}_{n,i}^W = (1 - (\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1} P_{n,i}$ and $\bar{a}_{n,i}^{(Z,W)} = (1 - (\mathbf{H}_{(\mathbf{Z}_n, \mathbf{W}_n)})_{ii})^{-1} P_{n,i}$

their expectations conditional on $\mathcal{Z}_n$. Note that since $0 \leq P_{n,i} \leq 1$, it follows that $\text{var}(P_{n,i} \mid \mathcal{Z}_n) \leq 1$. Then we can write:

$$\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n - \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n = \sum_i \sum_{j \neq i} \left(A_{n,i}^{(\mathbf{Z}_n,\mathbf{W}_n)} (\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii} T_j - \bar{a}_{n,i}^{(\mathbf{Z}_n,\mathbf{W}_n)} (\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii} P_j \right) + \sum_i \sum_{j \neq i} \left(A_{n,i}^{(\mathbf{W}_n)} (\mathbf{H}_{(\mathbf{W}_n)})_{ii} T_j - \bar{a}_{n,i}^{(\mathbf{W}_n)} (\mathbf{H}_{(\mathbf{W}_n)})_{ii} P_j \right).$$

Therefore, obtain that

$$\begin{aligned}
&\mathbb{E}[(\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n - \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n)^2/n^2 \mid \mathcal{Z}_n] \\
&\leq \mathbb{E}\left[\left(\sum_i \sum_{j \neq i} \left(A_{n,i}^{(\mathbf{Z}_n,\mathbf{W}_n)} (\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii} T_j - \bar{a}_{n,i}^{(\mathbf{Z}_n,\mathbf{W}_n)} (\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii} P_j \right) / n \right)^2 \,\middle|\, \mathcal{Z}_n \right] \\
&\quad + \mathbb{E}\left[\left(\sum_i \sum_{j \neq i} \left(A_{n,i}^{(\mathbf{W}_n)} (\mathbf{H}_{(\mathbf{W}_n)})_{ii} T_j - \bar{a}_{n,i}^{(\mathbf{W}_n)} (\mathbf{H}_{(\mathbf{W}_n)})_{ii} P_j \right) / n \right)^2 \,\middle|\, \mathcal{Z}_n \right] \\
&\leq \frac{C}{n^2} \left((K+L) \frac{1}{(1-C_2)^2} + \frac{2}{(1-C_2)^2} \mathbf{P}_n'\mathbf{P}_n \right) + \frac{C}{n^2} \left(L \frac{1}{(1-C_2)^2} + \frac{2}{(1-C_2)^2} \mathbf{P}_n'\mathbf{P}_n \right),
\end{aligned}$$

where the first line follows from triangle inequality, and the second line follows from applying Lemma A.3 with $\mathcal{F} = \mathcal{Z}_n$, and the implication of Assumption R that

$$\sup_n \sup_{i \leq n} \text{var}\left(A_{n,i}^{(Z,W)} \mid \mathcal{Z}_n \right) \leq \frac{1}{(1-C_2)^2}, \qquad \sup_n \sup_{i \leq n} \text{var}\left(A_{n,i}^{(W)} \mid \mathcal{Z}_n \right) \leq \frac{1}{(1-C_2)^2},$$

$$\sum_{i=1}^n A_{n,i}^W A_{n,i}^W \leq \frac{1}{1-C_2} \sum_{i=1}^n P_{n,i}^2, \qquad \sum_{i=1}^n A_{n,i}^{(Z,W)} A_{n,i}^{(Z,W)} \leq \frac{1}{1-C_2} \sum_{i=1}^n P_{n,i}^2.$$

Next, by Assumption R, we can apply the law of large numbers given in Lemma A.4 to $n^{-1}\sum_{i=1}^n P_{n,i}^2$ to get $n^{-1}\sum_{i=1}^n P_{n,i}^2 = \mathbb{E}[P^2] + o_p(1) = O_p(1)$. Also, $(K+L)/n^2 = o(1)$ by Assumption MI, so that

$$\mathbb{E}[(\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n - \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n)^2/n^2 \mid \mathcal{Z}_n] \leq o_p(1).$$

Therefore, by Markov inequality and the dominated convergence theorem,

$$\mathbf{T}_n'\mathbf{G}_n\mathbf{T}_n/n = \mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n + o_p(1),$$

which proves assertion (i). □

Lemma A.6. *Suppose Assumptions R and MI hold. Let (Y, T, P, R) be distributed according to the limiting distribution $F^{Y,T,R,P}$. Then*

(i) $\mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n = \mathbb{E}[T(P - \mathbb{E}[P \mid X])] + o_p(1)$; *and*

(ii) $\mathbf{R}_n'\mathbf{G}_n\mathbf{P}_n/n = \mathbb{E}[Y(P - \mathbb{E}[P \mid X])] + o_p(1)$,

where $\mathbf{G}_n$ is defined in Equation (A.1).

Proof. Again, I will only prove Part (i), Part (ii) follows by similar arguments. To this end, write $\mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n$ as

$$\mathbf{P}_n'\mathbf{G}_n\mathbf{P}_n/n = \mathbf{P}_n'(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)})\mathbf{P}_n/n - \mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n.$$

I will prove the assertion in two steps. First, I will prove that

$$\mathbf{P}_n'(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)})\mathbf{P}_n/n = \mathbb{E}[TP] + o_p(1). \tag{A.4}$$

Second, I will prove that

$$\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n = \mathbb{E}[T\mathbb{E}[P \mid X]] + o_p(1). \tag{A.5}$$

Combining (A.4) with (A.5) then yields the result.

To prove (A.4), note that

$$\begin{aligned}
\|\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n/\sqrt{n}\|_F^2 &= \|\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{P}_n - \mathbf{P}_n^L)/\sqrt{n}\|_F^2 \\
&= \mathrm{tr}((\mathbf{P}_n - \mathbf{P}_n^L)(\mathbf{P}_n - \mathbf{P}_n^L)'/n) - \mathrm{tr}(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{P}_n - \mathbf{P}_n^L)(\mathbf{P}_n - \mathbf{P}_n^L)'/n) \\
&\leq \mathrm{tr}((\mathbf{P}_n - \mathbf{P}_n^L)(\mathbf{P}_n - \mathbf{P}_n^L)'/n) = \sum_{i=1}^n (P_{n,i}^L - P_{n,i})^2/n \to 0 \qquad \text{a.s.},
\end{aligned} \tag{A.6}$$

where the first equality follows from $\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n^L = 0$, the second equality follows from the definition of Frobenius norm, and the last line follows from the fact that $\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)}$ is positive semi-definite so that $\mathrm{tr}(\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)}(\mathbf{P}_n - \mathbf{P}_n^L)(\mathbf{P}_n - \mathbf{P}_n^L)'/n) \geq 0$ and Assumption MI. Therefore, we obtain

$$\begin{aligned}
\|\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n/n\|_F &\leq \|\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\sqrt{n}\|_F \|\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n/\sqrt{n}\|_F \\
&= \left(n^{-1}\sum_i \frac{P_{n,i}^2}{(1 - (\mathbf{H}_{(\mathbf{Z}_n,\mathbf{W}_n)})_{ii})^2}\right)^{1/2} \|\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)}\mathbf{P}_n/\sqrt{n}\|_F \\
&\leq \frac{1}{1 - C_2}\left(n^{-1}\sum_i P_{n,i}^2\right)^{1/2} o_p(1) \\
&= o_p(1),
\end{aligned} \tag{A.7}$$

where the first line follows since the Frobenius norm is sub-multiplicative, the second line follows from the definition of Frobenius norm, the third line follows from the result (A.6) and Assumption R, and the last line follows from applying the law of large numbers given in Lemma A.4 to $n^{-1}\sum_{i=1}^n P_{n,i}^2$. Therefore, we obtain

$$\mathbf{P}_n'(\mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{(\mathbf{Z}_n,\mathbf{W}_n)})^{-1}\mathbf{M}_{(\mathbf{Z}_n,\mathbf{W}_n)})\mathbf{P}_n/n = \mathbf{P}_n'\mathbf{P}_n/n + o_p(1).$$

Since by Assumption R and Lemma A.4, $n^{-1}\sum_{i=1}^n P_{n,i}^2 \to \mathbb{E}[P^2] = \mathbb{E}[TP]$, assertion (A.4) follows.

Now I prove assertion (A.5). Let $A_{n,i} = (1 - (\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1}P_{n,i}$, and denote by $\bar{a}_{n,i}^W = (1 - (\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1}P_{n,i}^X$ its expectation conditional on $\{X_{n,i}\}_{i=1}^n$, where $P_{n,i}^X = \mathbb{E}[P_{n,i} \mid X_{n,i}]$. Note that since $0 \leq P_{n,i} \leq 1$, it follows that

$\text{var}(P_{n,i} \mid X_{n,i}) \leq 1$. Therefore, applying Lemma A.3 with $\mathcal{F} = \{X_{n,i}\}_{i=1}^n$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\sum_i \sum_{j \neq i} \left(A_{n,i}(\mathbf{H}_{\mathbf{W}_n})_{ij} P_j - a_{n,i}(\mathbf{H}_{\mathbf{W}_n})_{ij} \mathbb{E}[P_j \mid W_j]\right)/n\right)^2 \,\middle|\, \{X_{n,i}\}_{i=1}^n\right] \\
&\leq \frac{C}{n^2}\left(L\frac{1}{(1-C_2)^2} + \frac{2}{(1-C_2)^2}\sum_i (\mathbb{E}[P_{n,i} \mid X_{n,i}])^2\right) \\
&\leq \frac{C}{n^2}\left(L\frac{1}{(1-C_2)^2} + \frac{2n}{(1-C_2)^2}\right) \\
&= o_p(1),
\end{aligned}$$

where the first line follows from the implication of Assumption R, $(1 - (\mathbf{H}_{\mathbf{W}_n})_{ii})^{-1} \leq 1/(1 - C_2)$, the second line follows from $|P_{n,i}| \leq 1$, and the last line follows from $L \leq n$. It therefore follows by Markov inequality and the dominated convergence theorem that

$$\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n = (\mathbf{P}_n^X)'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n^X/n + o_p(1)$$

where $\mathbf{P}_n^X$ is an n-vector with ith element given by $P_{n,i}^X$. Let $P_{n,i}^{X,L} = \mathbb{E}^*[P_{n,i} \mid W_{n,i}]$. Now, by arguments as in Equations (A.6) and (A.7) with $\mathbf{P}_n$ replaced by $\mathbf{P}_n^X$ and $\mathbf{P}_n^L$ replaced by $\mathbf{P}_n^{X,L}$, we have that:

$$\|(\mathbf{P}_n^X)'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}\mathbf{M}_{\mathbf{W}_n}\mathbf{P}_n^X/n\|_F = o_p(1).$$

Since $(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n}) = \mathbf{I}_n - (\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}\mathbf{M}_{\mathbf{W}_n}$, it follows that

$$\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n = n^{-1}\sum_i \mathbb{E}[P_{n,i} \mid X_{n,i}]\mathbb{E}[P_{n,i} \mid X_{n,i}] + o_p(1).$$

By Assumption R, we can apply Lemma A.4 to obtain

$$\begin{aligned}
\mathbf{P}_n'(\mathbf{I}_n - \mathbf{D}_{\mathbf{W}_n})^{-1}(\mathbf{H}_{\mathbf{W}_n} - \mathbf{D}_{\mathbf{W}_n})\mathbf{P}_n/n &= \mathbb{E}\left[\mathbb{E}[P \mid X]^2\right] + o_p(1) \\
&= \mathbb{E}\left[T\mathbb{E}[P \mid X]\right] + o_p(1),
\end{aligned}$$

which prove assertion (A.5). □

A.2 Proofs

Proof of Lemma 1.1. First consider part (i). Since $\mathbb{E}[\tilde{P}_i^L T_i] \neq 0$, it follows by the continuous mapping theorem that $\hat{\beta}_{\hat{\mathbf{P}}} \xrightarrow{p} \mathbb{E}[\tilde{P}_i^L Y_i]/\mathbb{E}[\tilde{P}_i^L T_i]$. Part (i) then follows by Lemma A.1.

Next, consider Part (ii). The reverse two-stage least squares estimator can be written as:

$$\hat{\beta}_{\text{RTSLS}} = \frac{\hat{\pi}_1'(\mathbf{Z}_\perp'\mathbf{Z}_\perp/n)\hat{\pi}_1}{\hat{\pi}_1'(\mathbf{Z}_\perp'\mathbf{Z}_\perp/n)\hat{\pi}_2},$$

where $\hat{\pi}_1$ and $\hat{\pi}_2$ are least-squares estimators of π_1 and π_2 that are based on Equations (1.3)–(1.4). Since $\mathbb{E}[(Z_i, W_i)(Z_i, W_i)']$ is full rank, and the data is iid with finite second moments, these least-squares estimators are consistent for π_1 and π_2. Also, by the law of large numbers, the fact that $\mathbb{E}[W_i W_i']$ is full rank, and the continuous mapping theorem:

$$\begin{aligned}\mathbf{Z}_\perp'\mathbf{Z}_\perp/n &= \mathbf{Z}'\mathbf{Z}/n - (\mathbf{Z}'\mathbf{W}/n)(\mathbf{W}'\mathbf{W}/n)^{-1}(\mathbf{W}'\mathbf{Z}/n) \\ &\xrightarrow{p} \mathbb{E}[Z_i Z_i'] - \mathbb{E}[Z_i W_i'](\mathbb{E}[W_i W_i'])^{-1}\mathbb{E}[W_i Z_i'] = \mathbb{E}[\tilde{Z}_i \tilde{Z}_i'].\end{aligned}$$

Hence

$$\hat{\beta}_{\text{RTSLS}} \xrightarrow{p} \frac{\pi_1 \mathbb{E}[\tilde{Z}_i \tilde{Z}_i'] \pi_1}{\pi_1 \mathbb{E}[\tilde{Z}_i \tilde{Z}_i'] \pi_2},$$

which, combined with Lemma A.1, proves the assertion.

Finally, too prove Part (iii), it suffices to show that

$$\min \text{eig}(\hat{S}^{-1}\hat{\Xi}) \xrightarrow{p} \min \text{eig}(S^{-1}\Xi), \tag{A.8}$$

since then $\hat{\beta}_{\hat{S},\hat{\Xi}} \xrightarrow{p} \beta_S$ by the continuous mapping theorem. To show (A.8), note that $\min \text{eig}(\hat{S}^{-1}\hat{\Xi})$ is the minimum of the function

$$\hat{\mathcal{D}}_{\mathcal{S}}(\omega) = \frac{\omega'\hat{\Xi}\omega}{\omega'\hat{S}\omega}, \qquad\qquad \omega \in \mathcal{S}^1,$$

where $\mathcal{S}^1$ denotes the unit circle in $\mathbb{R}^2$, a compact space. Therefore, if $\hat{\mathcal{D}}_{\mathcal{S}}(\omega)$ converges uniformly to the limiting function $\mathcal{D}_{\mathcal{S}}(\omega) = \omega'\Xi\omega/(\omega' S\omega)$, then $\min_\omega \hat{\mathcal{D}}_{\mathcal{S}}(\omega) \xrightarrow{p} \min_\omega \mathcal{D}_{\mathcal{S}}(\omega)$ by standard arguments (see, for example Newey and McFadden, 1994). To prove uniform convergence, I will use the arguments in Chao and Swanson (2005). Fix some $\omega \in \mathcal{S}^1$, and note that:

$$\begin{aligned}|\hat{\mathcal{D}}_{\mathcal{S}}(\omega) - \mathcal{D}_{\mathcal{S}}(\omega)| &= \left|\frac{\omega'\hat{\Xi}\omega}{\omega'\hat{S}\omega} - \mathcal{D}_{\mathcal{S}}(\omega)\frac{\omega'\hat{S}\omega}{\omega'\hat{S}\omega}\right| = \frac{1}{|\omega'\hat{S}\omega|}\left|\omega'\hat{\Xi}\omega - \mathcal{D}_{\mathcal{S}}(\omega)\omega'\hat{S}\omega\right| \\ &= \frac{1}{|\omega'\hat{S}\omega|}\left|\omega'(\hat{\Xi} - \Xi)\omega - \mathcal{D}_{\mathcal{S}}(\omega)\omega'(\hat{S} - S)\omega\right| \\ &\leq \frac{1}{|\omega'\hat{S}\omega|}\left(\left|\omega'(\hat{\Xi} - \Xi)\omega\right| + \mathcal{D}_{\mathcal{S}}(\omega)\left|\omega'(\hat{S} - S)\omega\right|\right),\end{aligned}$$

where the first line follows from the definition of $\hat{\mathcal{D}}_{\hat{\mathcal{S}}}$, the second line follows from the definition of $\mathcal{D}_{\hat{\mathcal{S}}}$, and the third line follows by triangle inequality. I now bound all three terms in the last the expression uniformly in ω. Since the trace operator is an inner product under Frobenius norm $\|A\|_F = \sqrt{\text{tr}(AA')}$, by Cauchy-Schwarz inequality:

$$\begin{aligned}|\omega'(\hat{\Xi} - \Xi)\omega| &= \left|\text{tr}\left(\omega\omega'(\hat{\Xi} - \Xi)\right)\right| \leq \sqrt{\text{tr}(\omega\omega'\omega\omega')}\|\hat{\Xi} - \Xi\|_F \\ &= \|\hat{\Xi} - \Xi\|_F = o_p(1),\end{aligned}$$

where the second line follows from $\omega'\omega = 1$ since $\omega \in \mathcal{S}^1$ and $\hat{\Xi} \xrightarrow{p} \Xi$ so that $\|\hat{\Xi} - \Xi\|_F = o_p(1)$. By an identical

argument, we also have $|\omega'(\hat{S} - S)\omega| = o_p(1)$. Finally, to bound $1/|\omega'\hat{S}\omega|$, note that since $\hat{S} \xrightarrow{p} S > 0$, $\omega'\hat{S}\omega > 0$ with probability approaching 1, so that $1/|\omega'\hat{S}\omega| < C$ for some $C < \infty$ with probability approaching 1. Hence:

$$|\hat{\mathcal{D}}_{\mathcal{S}}(\omega) - \mathcal{D}_{\mathcal{S}}(\omega)| \leq o_p(1) + o_p(1)\mathcal{D}_{\mathcal{S}}(\omega),$$

since $\mathcal{D}_{\mathcal{S}}(\omega)$ is bounded by $\max \text{eig}(S^{-1}\Xi)$, it follows that $\sup_\omega |\hat{\mathcal{D}}_{\mathcal{S}}(\omega) - \mathcal{D}_{\mathcal{S}}(\omega)| = o_p(1)$ as required. □

Proof of Theorem 1.1. By Lemma 1.1, $\Xi_{12} = \mathbb{E}[Y_i \tilde{P}_i^L]$ and $\Xi_{22} = \mathbb{E}[T_i \tilde{P}_i^L]$. Since by Assumption L, $\mathbb{E}[\tilde{P}_i^L \mid X_i] = 0$, we can apply Lemma A.2 with $A_i = \tilde{P}_i^L$ to get

$$\frac{\Xi_{12}}{\Xi_{22}} = \frac{\int \sum_{j=1}^{J_x - 1} \alpha(p_{j,x}; x)(p_{j,x} - p_{j-1,x}) \mathbb{E}[\tilde{P}_i^L \mid X_i = x, P_i > p_{j,x}] \mathbb{P}(P_i > p_{j,x} \mid X_i = x) \, \mathrm{d}F^X(x)}{\int \sum_{j=1}^{J_x - 1} (p_{j,x} - p_{j-1,x}) \mathbb{E}[\tilde{P}_i^L \mid X_i = x, P_i > p_{j,x}] \mathbb{P}(P_i > p_{j,x} \mid X_i = x) \, \mathrm{d}F^X(x)},$$

which yields the result for Ξ_{12}/Ξ_{22}.

Second, by Lemma 1.1, $\Xi_{12} = \mathbb{E}[T_i \tilde{R}_i^L]$ and $\Xi_{11} = \mathbb{E}[Y_i \tilde{R}_i^L]$. Since by Assumption L, $\mathbb{E}[\tilde{R}_i^L \mid X_i] = 0$, applying Lemma A.2 with $A_i = \tilde{R}_i^L$ yields the result for Ξ_{11}/Ξ_{12}. □

Proof of Corollary 1.1. Let $\mathbb{P}(P_i = p_{j,x} \mid X_i = x) = s_{j,x}$. If the linear approximations (1.3)–(1.4) are exact, then $P_i = P_i^L$ and $R_i = R_i^L$. We can therefore write:

$$\mathbb{E}[R_i^L \mid P_i = p_{j,x}, X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] + \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}),$$

so that

$$\mathbb{E}[R_i^L \mid X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] + \sum_{j'=1}^{J_x - 1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m=j'+1}^{J_x} s_{m,x},$$

and

$$\begin{aligned}\mathbb{E}[\tilde{R}_i^L \mid P_i > p_{j,x}, X_i = x] = \mathbb{E}[R_i^L \mid P_i = p_{1,x}, X_i = x] &+ \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) + \\ &+ \frac{1}{\sum_{m=j+1}^{J_x} s_{m,x}} \sum_{j'=j}^{J_x - 1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m=j'+1}^{J_x} s_{m,x}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[\tilde{R}_i^L \mid P_i > p_{j,x}, X_i = x] \mathbb{P}(P_i > p_{j,x} \mid X_i = x) = &\sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m'=j+1}^{J_x} s_{m',x} \sum_{m=1}^{j'} s_{m,x} \\ &+ \sum_{j'=j}^{J_x - 1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m'=1}^{j} s_{m',x} \sum_{m=j'+1}^{J_x} s_{m,x}.\end{aligned}$$

By Theorem 1.1, we therefore have:

$$\zeta_j(x) = (p_{j+1,x} - p_{j,x}) \sum_{j'=1}^{j-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m'=j+1}^{J_x} s_{m',x} \sum_{m=1}^{j'} s_{m,x} \\ + (p_{j+1,x} - p_{j,x}) \sum_{j'=j}^{J_x-1} \alpha(p_{j',x}; x)(p_{j'+1,x} - p_{j',x}) \sum_{m'=1}^{j} s_{m',x} \sum_{m=j'+1}^{J_x} s_{m,x}.$$

If $\alpha(p_{j,x}; x) \geq 0$ for all j and x, then all the terms in this expression are non-negative, so that $\zeta_j(x)$ is non-negative.

To obtain the expressions for $\zeta_1(x)$ and $\theta_1(x)$ in the special case that $J_x = 2$, note that since $\mathbb{E}[P_i \mid X_i = x] = s_{1,x}p_{1,x} + s_{2,x}p_{2,x}$, we have

$$\begin{aligned}\theta_1(x) &= (p_{2,x} - p_{1,x})[p_{2,x} - \mathbb{E}[P_i \mid X_i = x]]s_{2,x} \\ &= (p_{2,x} - p_{1,x})[p_{2,x}(1 - s_{2,x}) - p_{1,x}s_{1,x}]s_{2,x} = (p_{2,x} - p_{1,x})^2 s_{1,x}s_{2,x}.\end{aligned}$$

On the other hand,

$$\operatorname{var}(P_i \mid X_i = x) = (p_{2,x} - \mathbb{E}[P_i \mid X_i = x])^2 s_{2,x} + (p_{1,x} - \mathbb{E}[P_i \mid X_i = x])^2 s_{1,x} = (p_{2,x} - p_{1,x})^2 s_{1,x}s_{2,x}$$

Secondly, since $R_i^L = R_i$,

$$\begin{aligned}\mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] &= \mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] \\ \mathbb{E}[Y_i \mid X_i = x, P_i = p_{2,x}] &= \mathbb{E}[Y_i \mid X_i = x, P_i = p_{1,x}] + \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x}),\end{aligned}$$

so that

$$\begin{aligned}\mathbb{E}[\tilde{R}_i^L \mid X_i = x, P_i > p_{1,x}] &= \mathbb{E}[\tilde{R}_i^L \mid X_i = x, P_i = p_{2,x}] \\ &= \mathbb{E}[Y_i \mid X_i = x, P_i = p_{0,x}] + \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x}) - \mathbb{E}[Y_i \mid X_i = x] \\ &= \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x})(1 - s_{2,x}) \\ &= \alpha(p_{1,x}; x)(p_{2,x} - p_{1,x})s_{1,x}.\end{aligned}$$

Therefore, it follows that:

$$\zeta_1(x) = (p_{2,x} - p_{1,x})^2 s_{1,x}s_{2,x}\alpha(p_{1,x}; x),$$

which completes the proof. □

Proof of Theorem 1.2. Using the matrix notation from Equation (A.1),

$$\hat{\beta}_{\text{UJIVE}} = \frac{\mathbf{Y}_n'\mathbf{G}_n'\mathbf{T}_n/n}{\mathbf{T}_n\mathbf{G}_n'\mathbf{T}_n/n}$$

By Lemma A.5 and Lemma A.6,

$$\mathbf{Y}_n'\mathbf{G}_n'\mathbf{T}_n/n = \mathbb{E}[Y(P - \mathbb{E}[P \mid X])] + o_p(1), \qquad \mathbf{T}_n\mathbf{G}_n'\mathbf{T}_n/n = \mathbb{E}[T(P - \mathbb{E}[P \mid X])] + o_p(1).$$

Next, since the limiting distribution of the data satisfies Assumption IV (iii), $\mathbb{E}[T(P - \mathbb{E}[P \mid X])] > 0$, so that by the continuous mapping theorem,

$$\frac{\mathbf{T}_n' \mathbf{G}_n \mathbf{Y}_n}{\mathbf{T}_n' \mathbf{G}_n \mathbf{T}_n} = \frac{\mathbb{E}[Y(P - \mathbb{E}[P \mid X])]}{\mathbb{E}[T(P - \mathbb{E}[P \mid X])]} + o_p(1).$$

The assertion of the Theorem then follows by applying Lemma A.2 with $A_i = P - \mathbb{E}[P \mid X]$. □

Appendix B

Appendix to Chapter 2

We first define some additional notation. Write the reduced-form based on Equations (1.5) as:

$$\begin{pmatrix} Y_i & X_i \end{pmatrix} = \begin{pmatrix} Z_i & W_i \end{pmatrix}' \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix} + V_i',$$

where $\pi_{11} = \gamma + \pi_{12}\beta$ and $\pi_{21} = \delta + \pi_{22}\beta$, and $V_i = (\epsilon_i + \nu_i\beta, \nu_i)'$, and let $\mathbf{V}$ be the $N \times 2$ matrix with ith row equal to V_i'. Denote the upper $K_N \times 2$ submatrix of the matrix of reduced-form coefficients by $\Pi_1 = (\pi_{11}, \pi_{12})$. Let:

$$\Gamma = \begin{pmatrix} 1 & 0 \\ -\beta & 1 \end{pmatrix}.$$

Let $\Omega = \mathbb{E}[V_i V_i']$ denote the reduced-form covariance matrix. Then:

$$\Omega = \Gamma^{-1\prime}\Sigma\Gamma^{-1} = \begin{pmatrix} \Sigma_{11} + 2\Sigma_{12}\beta + \Sigma_{22}\beta^2 & \Sigma_{12} + \Sigma_{22}\beta \\ \Sigma_{21} + \Sigma_{22}\beta & \Sigma_{22} \end{pmatrix}.$$

Let $\mathcal{W}_d(f, V, V^{-1}M)$ denote a d-dimensional non-central Wishart distribution with f degrees of freedom, scale parameter V, and non-centrality parameter M. Let $\mathbf{S}^{1/2}$ denote the symmetric square root of a symmetric positive semi-definite matrix $\mathbf{S}$.

B.1 Auxilliary Lemmata

Lemma B.1 (Lemmata 1 and 2, Bekker, 1994). *Consider the quadratic form*

$$Q = (M + U)'C(M + U),$$

where $M \in \mathbb{R}^{N\times S}, C \in \mathbb{R}^{N\times N}$ are non-stochastic, C is symmetric and idempotent with rank J_N which may depend on N, and $U = (u_1, \ldots, u_N)'$, with $u_i \sim [0, \Omega]$ iid. Let $a \in \mathbb{R}^S$ be a non-stochastic vector. Then:

(i) *If u_i has finite fourth moments:*

$$\begin{aligned}
\mathbb{E}[Q \mid C] &= M'CM + J_N\Omega, \\
\operatorname{var}(Qa \mid C) &= a'\Omega aM'CM + a'M'CMa\Omega + \Omega aa'M'CM + MCMaa'\Omega \\
&\quad + J_N(a'\Omega a\Omega + \Omega aa'\Omega) \\
&\quad + d_C'd_C[\mathbb{E}(a'u)^2uu' - a'\Omega aa'\Omega - a'\Omega a\Omega] + 2d_C'CMa\mathbb{E}[(a'u)uu'] \\
&\quad + M'Cd_C\mathbb{E}[(a'u)^2u'] + \mathbb{E}[(a'u)^2u]d_C'CM,
\end{aligned}$$

where $d_C = \operatorname{diag}(C)$. If the distribution of u_i is Normal, the last two lines of the variance expression equals zero.

(ii) *Suppose that the distribution of u_i is Normal, and that, as $N \to \infty$:*

$$M'CM/N \to Q_{CM}, \qquad J_N/N \to \alpha_r,$$

where the elements c_{is} of C may depend on N. Then:

$$\sqrt{N}\left(Qa/N - \mathbb{E}Qa/N\right) \Rightarrow \mathcal{N}\left(0, V\right),$$

where

$$V = a'\Omega aQ_{CM} + a'Q_{CM}a\Omega + \Omega aa'Q_{CM} + Q_{CM}aa'\Omega + \alpha_r(a'\Omega a\Omega + \Omega aa'\Omega).$$

Lemma B.2. *Consider a sequence of random matrices* $\{X_N\}_{N=1}^{\infty}$ *with* $X_N \sim \mathcal{W}_S(J_N, \Omega, \Omega^{-1}\Xi_N)$. *Suppose that* $\Xi_N/N \to \Xi$, *and that* $J_N/N = \alpha + o(N^{-1/2})$, $\alpha > 0$. *Then, for any vector* $a \in \mathbb{R}^S$

$$N^{-1/2}\left(X_N a/N - (\Xi_N/N + \alpha\Omega)a\right) \Rightarrow \mathcal{N}\left(0, (a'\Omega a\Xi + a'\Xi a\Omega + \Omega aa'\Xi + \Xi aa'\Omega) + \alpha(a'\Omega a\Omega + \Omega aa'\Omega)\right).$$

Proof. By definition of a non-central Wishart distribution, we can decompose $X_N = (U + M)'(U + M)$, where $U = (u_1, \ldots, u_{J_N})'$, $u_j \sim N(0, \Omega)$ iid, $M'M = \Xi_N$, and $\Xi_N/J_N \to \Xi/\alpha$. Hence, we can apply Lemma B.1 (ii) with $C = \mathbf{I}_{J_N}$ to get:

$$J_N^{-1/2}(X_N a - (\Xi_N + J_N\Omega)a) \Rightarrow \mathcal{N}\left(0, \alpha^{-1}(a'\Omega a\Xi + a'\Xi a\Omega + \Omega aa'\Xi + \Xi aa'\Omega) + a'\Omega a\Omega + \Omega aa'\Omega\right),$$

which yields the result. □

Lemma B.3. *Suppose Assumptions 1, 2(i), 3 and 4 hold. Then:*

$$\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N \xrightarrow{p} \Psi + (1 - \alpha_L)\Omega, \tag{B.1a}$$

$$\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N \xrightarrow{p} \Psi + \alpha_K\Omega, \tag{B.1b}$$

where

$$\Psi = \begin{pmatrix} \Lambda_{11} + 2\Lambda_{12}\beta + \Lambda_{22}\beta^2 & \Lambda_{12} + \Lambda_{22}\beta \\ \Lambda_{12} + \Lambda_{22}\beta & \Lambda_{22}. \end{pmatrix} \tag{B.2}$$

These probability limits also hold conditional on $\overline{\mathbf{Z}}$.

Proof. First we establish the probability limit of $\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N$. By Lemma B.1 (i):

$$\mathbb{E}[\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N \mid \mathbf{Z}_\perp] = (K_N/N)\Omega. \tag{B.3}$$

Fix $a \in \mathbb{R}^2$. Since $\mathbf{P}_{\mathbf{Z}_\perp}$ is a projection matrix, $0 \leq (\mathbf{P}_{\mathbf{Z}_\perp})_{ii} \leq 1$. Hence, $\sum_i (\mathbf{P}_{\mathbf{Z}_\perp})_{ii}^2 \leq$

$\sum_i (\mathbf{P}_{\mathbf{Z}_\perp})_{ii} \leq K_N$. Therefore:

$$\begin{aligned}
\operatorname{var}(\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}a/N) &= \mathbb{E}\operatorname{var}(\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}a/N \mid \mathbf{P}_{\mathbf{Z}_\perp}) \\
&= \mathbb{E}\left[\operatorname{tr}(\mathbf{P}_{\mathbf{Z}_\perp}/N^2)\right](a'\Omega a\Omega + \Omega aa'\Omega) \\
&\quad + \mathbb{E}\left[N^{-2}\textstyle\sum_i (\mathbf{P}_{\mathbf{Z}_\perp})_{ii}^2\right][\mathbb{E}(a'V_i)^2 V_i V_i' - a'\Omega aa'\Omega - a'\Omega a\Omega] \qquad \text{(B.4)} \\
&\leq \frac{K_N}{N^2}(a'\Omega a\Omega + \Omega aa'\Omega) + \frac{K_N}{N^2}[\mathbb{E}(a'v_i)^2 v_i v_i' - a'\Omega aa'\Omega - a'\Omega a\Omega] \\
&= O(K_N/N^2).
\end{aligned}$$

Combining Equations (B.3) and (B.4) with Assumption 3 yields :

$$\mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N \xrightarrow{p} \alpha_K\Omega. \qquad \text{(B.5)}$$

By similar arguments:

$$\mathbf{V}'\mathbf{M}_{\mathbf{W}}\mathbf{V}/N \xrightarrow{p} (1-\alpha_L)\Omega. \qquad \text{(B.6)}$$

Next, by Assumption 2 (i), $\mathbb{E}[\Pi_1'\mathbf{Z}_\perp'\mathbf{V}/N \mid \mathbf{Z}_\perp] = 0$, so that:

$$\begin{aligned}
\operatorname{var}\left(\Pi_1'\mathbf{Z}_\perp'\mathbf{V}a/N\right) &= \mathbb{E}\left[\operatorname{var}\left(\Pi_1'\mathbf{Z}_\perp'\mathbf{V}a/N \mid \mathbf{Z}_\perp\right)\right] = (a'\Omega a)\mathbb{E}\left[\Pi_1'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi_1/N^2\right] \\
&= (a'\Omega a)\Gamma^{-1'}\mathbb{E}\left[\Lambda_N/N^2\right]\Gamma^{-1} = O(1/N),
\end{aligned}$$

where the last equality follows by Assumption 4. Consequently:

$$\Pi_1\mathbf{Z}_\perp'\mathbf{V}/N \xrightarrow{p} 0. \qquad \text{(B.7)}$$

Combining the representation $\mathbf{Y}_\perp = \mathbf{Z}_\perp\Pi_1 + \mathbf{V}_\perp$ with the limits in Equations (B.6) and (B.7), and Assumption 4 establishes (B.1a):

$$\begin{aligned}
\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N &= \Pi_1'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi_1/N + \Pi_1'\mathbf{Z}_\perp\mathbf{V}/N + \mathbf{V}'\mathbf{Z}_\perp\Pi_1/N + \mathbf{V}'\mathbf{M}_{\mathbf{W}}\mathbf{V}/N \\
&= \Gamma^{-1}\Lambda_N\Gamma^{-1}/N + (1-\alpha_L)\Omega + o_p(1) \\
&= \Psi + (1-\alpha_L)\Omega.
\end{aligned}$$

Claim (B.1b) follows by similar arguments from Equations (B.5) and (B.7):

$$\begin{aligned}
\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N &= \Pi_1'\mathbf{Z}_\perp'\mathbf{Z}_\perp\Pi_1/N + \Pi_1'\mathbf{Z}_\perp\mathbf{V}/N + \mathbf{V}'\mathbf{Z}_\perp\Pi_1/N + \mathbf{V}'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{V}/N \\
&\xrightarrow{p} \Psi + \alpha_K\Omega.
\end{aligned}$$

This concludes the proof. □

B.2 Proofs of Theorems

Proof of Proof of Theorem 2.1. Combining Lemma B.3 with the condition $\hat{k} = k + o_p(1)$ yields:

$$\begin{aligned}(1-\hat{k})\overline{\mathbf{Y}}_\perp'\overline{\mathbf{Y}}_\perp/N + \hat{k}\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/N &= (1-k)(\Psi + (1-\alpha_L)\Omega) + k(\Psi + \alpha_K\Omega) + o_p(1) \\ &= \Psi + (1-\alpha_L - (1-\alpha_K-\alpha_L)k)\Omega + o_p(1).\end{aligned} \tag{B.8}$$

Since $\Sigma_{22} = \Omega_{22}$, the (2,2) element of (B.8) is given by:

$$(1-\hat{k})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{k}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N = \Lambda_{22} + (1-\alpha_L - (1-\alpha_K-\alpha_L)k)\Sigma_{22} + o_p(1).$$

By the condition on k, $\Lambda_{22} + (1-\alpha_L - (1-\alpha_K-\alpha_L)k)\Sigma_{22} > 0$, so that:

$$\left((1-\hat{k})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{k}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N\right)^{-1} = (\Lambda_{22} + (1-\alpha_L - (1-\alpha_K-\alpha_L)k)\Sigma_{22})^{-1} + o_p(1). \tag{B.9}$$

The (1,2) element in Equation (B.8) is given by:

$$\begin{aligned}&(1-\hat{k})\mathbf{X}_\perp'\mathbf{Y}_\perp/N + \hat{k}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{Y}_\perp/N = \Lambda_{12} + \Lambda_{22}\beta + (1-\alpha_L - (1-\alpha_K-\alpha_L)k)\Omega_{12} + o_p(1) \\ &= \Lambda_{12} + (1-\alpha_L - (1-\alpha_K-\alpha_L)k)\Sigma_{12} + (1-\alpha_L - (1-\alpha_K-\alpha_L)k)\Sigma_{22}\beta + \Lambda_{22}\beta + o_p(1).\end{aligned} \tag{B.10}$$

Applying Equations (B.9) and (B.10) to $\hat{\beta}_{\hat{k}}$:

$$\hat{\beta}_{\hat{k}} = \frac{(1-\hat{k})\mathbf{X}_\perp'\mathbf{Y}_\perp/N + \hat{k}x_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{Y}_\perp}{(1-\hat{k})\mathbf{X}_\perp'\mathbf{X}_\perp/N + \hat{k}\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\mathbf{X}_\perp/N} = \beta + \frac{\Lambda_{12} + ((1-k)(1-\alpha_L) + \alpha_K k)\Sigma_{12}}{\Lambda_{22} + ((1-k)(1-\alpha_L) + \alpha_K k)\Sigma_{22}} + o_p(1).$$

□

Proof of Proof of Corollary 2.1. The results for tsls, btsls and mbtsls follow directly from

Theorem 2.1. We therefore just need to derive the results for liml. Define

$$\hat{Q}_N(\phi) = \frac{\phi' \overline{\mathbf{Y}}_\perp' \overline{\mathbf{Y}}_\perp / N\phi}{\phi' \overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N\phi}.$$

Then

$$\hat{k}_{\text{LIML}} = \min_{\tilde{\beta}} \frac{(1, -\tilde{\beta}) \overline{\mathbf{Y}}_\perp' \overline{\mathbf{Y}}_\perp / N (1, -\tilde{\beta})'}{(1, -\tilde{\beta}) \overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N (1, -\tilde{\beta})'} = \min_{\phi \in S^1} \hat{Q}_N(\phi),$$

where S^1 denotes the unit circle in $\mathbb{R}^2$. Applying Lemma B.3 yields

$$\hat{Q}_N(\phi) \xrightarrow{p} \frac{\phi'(\Psi + (1 - \alpha_L)\Omega)\phi}{(1 - \alpha_L - \alpha_K)\phi' \Omega \phi} \equiv \frac{\phi' T \phi}{\phi' T_\perp \phi} \equiv Q(\phi),$$

where we define $T = \Psi + (1 - \alpha_L)\Omega$ and $T_\perp = (1 - \alpha_L - \alpha_K)\Omega$. Assumption 2 (i) guarantees that the denominator is non-zero for any value of ϕ. The minimum of $Q(\phi)$ is achieved at

$$\begin{aligned} \min_{\phi \in S^1} Q(\phi) &= \frac{1 - \alpha_L}{1 - \alpha_K - \alpha_L} + \frac{1}{1 - \alpha_L - \alpha_K} \min_{\phi \in S^1} \frac{\phi' \Psi \phi}{\phi' \Omega \phi} \\ &= \frac{1 - \alpha_L}{1 - \alpha_K - \alpha_L} + \frac{\min \text{eig}(\Sigma^{-1} \Lambda)}{1 - \alpha_K - \alpha_L} = k_{\text{LIML}}, \end{aligned}$$

where the last line follows since the eigenvalues of $\Omega^{-1}\Psi$ correspond to the eigenvalues of $\Sigma^{-1}\Lambda$. The minimand ϕ_{LIML} is given by the eigenvector corresponding to the smallest eigenvalue of the matrix

$$\frac{1}{1 - \alpha_K - \alpha_L} \Omega^{-1} \left(\Psi + (1 - \alpha_L)\Omega \right).$$

We now need to show that

$$\hat{k}_{\text{LIML}} - k_{\text{LIML}} = \min_{\phi \in S^1} \hat{Q}_N(\phi) - Q(\phi_{\text{LIML}}) \xrightarrow{p} 0. \tag{B.11}$$

To this end, we first show that the convergence of the objective function is uniform,

$$\sup_{\phi \in S^1} |\hat{Q}_N(\phi) - Q(\phi)| \xrightarrow{p} 0. \tag{B.12}$$

Fix $\phi \in S^1$. By the triangle inequality,

$$
\begin{aligned}
&|\hat{Q}_N(\phi) - Q(\phi)| \leq \\
&\qquad \leq \frac{1}{|\phi'\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N|} \left| \phi' \overline{\mathbf{Y}}_\perp' \overline{\mathbf{Y}}_\perp \phi / N - Q(\phi) \phi' \overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N \right| \\
&\quad = \frac{1}{|\phi'\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N|} \left| \phi' (\overline{\mathbf{Y}}_\perp' \overline{\mathbf{Y}}_\perp / N - T) \phi - Q(\phi) \phi' \left(\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp \right) \phi \right| \\
&\quad \leq \frac{1}{|\phi'\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N|} \left(\left| \phi' (\overline{\mathbf{Y}}_\perp' \overline{\mathbf{Y}}_\perp / N - T) \phi \right| + Q(\phi) \left| \phi' \left(\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp \right) \phi \right| \right).
\end{aligned}
\tag{B.13}
$$

We now need to bound all three terms in the expression uniformly in ϕ. Because the trace operator is the inner product under Frobenius norm, by Cauchy-Schwarz inequality,

$$
\begin{aligned}
|\phi' (\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp) \phi| &= \left| \operatorname{tr} \left((\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp) \phi \phi' \right) \right| \\
&\leq \sqrt{\operatorname{tr}((\phi\phi')^2)} \| (\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp) \|_F \\
&= \| (\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N - T_\perp \|_F \\
&= o_p(1),
\end{aligned}
$$

where the third line follows since $\|\phi\|_2 = 1$, and the last line follows since $\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N \xrightarrow{p} T_\perp$ by Lemma B.3. By similar argument,

$$
|\phi' (\overline{\mathbf{Y}}_\perp' \overline{\mathbf{Y}}_\perp / N - T) \phi| = o_p(1).
$$

Finally, we bound the denominator. Because $\overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp / N \xrightarrow{p} T_\perp > 0$, $\phi' \overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N > 0$ wpa1, so that wpa1 $1/(|\phi' \overline{\mathbf{Y}}_\perp' M_{\mathbf{Z}_\perp} \overline{\mathbf{Y}}_\perp \phi / N|) < C$ for some $C < \infty$. Applying these bounds and the fact that $Q(\phi)$ is bounded implies that the right-hand side in (B.13) is $o_p(1)$, which implies (B.12).

Next, denote the argmin of $\hat{Q}_N(\phi)$ by $\hat{\phi}$. Note that $\hat{k}_{\text{LIML}}$ and hence $\hat{\phi}$ exists wpa1. We

can now establish (B.11), using the uniform convergence result (B.12),

$$\begin{aligned} Q(\phi_{\text{LIML}}) \leq Q(\hat{\phi}) &= \hat{Q}_N(\hat{\phi}) + (Q(\hat{\phi}) - \hat{Q}_N(\hat{\phi})) \leq \hat{Q}(\phi_{\text{LIML}}) + (Q(\hat{\phi}) - \hat{Q}_n(\hat{\phi})) \\ &= Q(\phi_{\text{LIML}}) + (\hat{Q}_N(\phi_{\text{LIML}}) - Q(\phi_{\text{LIML}})) + (Q(\hat{\phi}) - \hat{Q}_N(\hat{\phi})) \\ &= Q(\phi_{\text{LIML}}) + o_p(1). \end{aligned}$$

The probability limit for liml then follows by Theorem 2.1. □

Proof of Proof of Theorem 2.2. Under Assumption 2, we have:

$$\sqrt{\alpha_K}\begin{pmatrix} (\mathbf{Z}_\perp'\mathbf{Z}_\perp)^{-1/2}\mathbf{Z}_\perp'\mathbf{Y} \\ (\mathbf{Z}_\perp'\mathbf{Z}_\perp)^{-1/2}\mathbf{Z}_\perp'\mathbf{X} \end{pmatrix} \Big| \, \overline{\mathbf{Z}} \sim \mathcal{N}\left(\begin{pmatrix} \tilde{\pi}_{12}\beta + \tilde{\gamma} \\ \tilde{\pi}_{12} \end{pmatrix}, \alpha_K\Omega \otimes \mathbf{I}_{K_N}\right),$$

$$\overline{\mathbf{Y}}_\perp'\mathbf{M}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp \mid \overline{\mathbf{Z}} \sim \mathcal{W}_2(N - K_N - L_N, \Omega).$$

Moreover, these two statistics are independent. Let $b = (1, -\beta)'$ and $a = (\beta, 1)$. Assumption 6 then implies that unconditionally,

$$\begin{aligned} &\overline{\mathbf{Y}}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp \sim \mathcal{W}_2(K_N, \Gamma^{-1\prime}\Xi\Gamma^{-1}/\alpha_K + \Omega, \left(\Gamma^{-1\prime}\Xi\Gamma^{-1}/\alpha_K + \Omega\right)^{-1} K_N aa'\mu_\pi^2/\alpha_K), \\ &\overline{\mathbf{Y}}_\perp'\mathbf{M}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp \sim \mathcal{W}_2(N - K_N - L_N, \Omega), \end{aligned}$$

with the independence property preserved. Applying Lemma B.2 then after some algebra yields:

$$N^{1/2}\left(\mathbf{X}_\perp'\mathbf{M}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp b/N - (1 - \alpha_K - \alpha_L)\Sigma_{12}\right) \Rightarrow \mathcal{N}\left(0, (1 - \alpha_K - \alpha_L)V_\Sigma\right), \tag{B.14a}$$

$$N^{1/2}\left(\mathbf{X}_\perp'\mathbf{P}_{\mathbf{Z}_\perp}\overline{\mathbf{Y}}_\perp/Nb - (\alpha_K\Sigma_{12})\right) \Rightarrow \mathcal{N}\left(0, \alpha_K V_\Sigma + V_\Xi\right), \tag{B.14b}$$

where

$$\begin{aligned} V_\Sigma &= \Sigma_{22}\Sigma_{11} + \Sigma_{12}^2, \\ V_\Xi &= \Lambda_{22}\Sigma_{11} + \Lambda_{11}\Sigma_{22} + \alpha_K^{-1}\Lambda_{22}\Lambda_{11}. \end{aligned}$$

Equations (B.14) imply that

$$N^{1/2}\left(\mathbf{X}_{\perp}'\mathbf{P}_{\mathbf{Z}_{\perp}}\overline{\mathbf{Y}}_{\perp}/N+(1-k_{\text{MBTSLS}})\mathbf{X}_{\perp}'\mathbf{M}_{\mathbf{Z}_{\perp}}\overline{\mathbf{Y}}_{\perp}/N\right)b\Rightarrow\mathcal{N}\left(0,V_{\Xi}+\frac{\alpha_K(1-\alpha_L)}{1-\alpha_K-\alpha_L}V_{\Sigma}\right).$$

Since by Lemma B.3, $(\mathbf{X}_{\perp}'\mathbf{P}_{\mathbf{Z}_{\perp}}\mathbf{X}_{\perp}N+(1-k_{\text{MBTSLS}})\mathbf{X}_{\perp}'\mathbf{M}_{\mathbf{Z}_{\perp}}\mathbf{X}_{\perp}/N)^{-1}\xrightarrow{p}\Lambda_{22}^{-1}+o_p(1)$, this yields the claim in the theorem. □

Appendix C

Appendix to Chapter 3

C.1 Definitions and identities

First I state a couple of simple identifies that are used throughout the appendix. Then, in Appendix C.2 I state and prove some auxiliary Lemmata that are helpful for proving the main results. In Appendix C.3, I derive the likelihood expressions in Equations (3.8), (3.9), (3.13), (3.17) and (3.25). The propositions and theorems stated in the text are derived in Appendix C.4.

For any symmetric matrix $\Omega \in \mathbb{R}^{2\times 2}$, and vectors $a = (\beta, 1)'$ and $b = (1, -\beta)'$, $\beta \in \mathbb{R}$:

$$Q_{\mathcal{S}}(\beta,\Omega) + Q_{\mathcal{T}}(\beta,\Omega) = \operatorname{tr}(\Omega^{-1}T) \tag{C.1a}$$

$$|\Omega|\, a\Omega^{-1}a = b'\Omega b \tag{C.1b}$$

$$\Omega b b'\Omega = (b'\Omega b)\Omega - |\Omega| aa' \tag{C.1c}$$

$$\frac{(e_1\Omega^{-1}e_1)^2}{a'\Omega^{-1}a} = e_1'\Omega^{-1}e_1 - \frac{1}{b'\Omega b} \tag{C.1d}$$

$$\Omega^{-1}T\Omega^{-1} = \operatorname{tr}(\Omega^{-1}T)\Omega^{-1} - \frac{1}{|\Omega|}\begin{pmatrix} T_{22} & -T_{12} \\ -T_{12} & T_{11} \end{pmatrix} \tag{C.1e}$$

$$\left|Y_\perp' Y_\perp\right| = n^2\,|T| + (n - k_n - \ell_n)^2\,|S| + n(n - k_n - \ell_n)\operatorname{tr}\left(S^{-1}T\right)|S| \tag{C.1f}$$

All equalities follow from simple algebra. Secondly, I use the following properties of the Kronecker product:

$$a \otimes b' = b' \otimes a = ab' \qquad \text{vec}(ACB) = (B' \otimes A)\,\text{vec}(C) \tag{C.2}$$

for some vectors $a, b \in \mathbb{R}^d$, and conformable matrices A, B, C.

Denote the duplication, elimination, and commutation matrices by D_d, L_d and K_d (see Magnus and Neudecker (1980) for definitions of these matrices). Let $N_d = (I_{d^2} + K_{dd})/2$ be the symmetrizer matrix. Then for arbitrary matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$ (Magnus and Neudecker, 1979, 1980):

$$K_{m1} = K_{1m} = I_m \qquad (B \otimes A)K_{qn} = K_{pm}(A \otimes B) \tag{C.3a}$$

$$K_d D_d = D_d \qquad D_d L_d N_d = N_d \tag{C.3b}$$

C.2 Auxiliary Lemmata

Lemma C.1. *Suppose $P \sim W_d(\nu_n, V, M_n)$. Then:*

(i) *[Magnus and Neudecker, 1979, Theorem 4.4]The mean and variance of P are given by:*

$$\mathbb{E}[P] = \nu_n V + M_n \qquad \text{var}(\text{vec}(P)) = 2N_d\left[\nu_n(V \otimes V) + V \otimes M_n + M_n \otimes V\right]$$

where N_d is the symmetrizer matrix.

(ii) *Suppose $M_n/n \to M$, and $\nu_n/n = \alpha + o(n^{-1/2})$ where $\alpha < 1$. Then, as $n \to \infty$*

$$\sqrt{n}\,\text{vec}\left(P/n - \mathbb{E}[P/n]\right) \Rightarrow \mathcal{N}_{d^2}\left(0, 2N_d[\alpha(V \otimes V) + V \otimes M + M \otimes V]\right)$$

Proof. To prove part (ii), decompose P as $P = \sum_{i=1}^{\nu_n} X_i X_i'$, where $X_i \sim \mathcal{N}_d(\mu_i, V)$ such that $M_n = \sum_i \mu_i \mu_i'$. Suppose first that $\alpha > 0$. Then it follows by the Central Limit Theorem that:

$$\nu_n^{-1/2}\,\text{vec}(P - \mathbb{E}[P]) \Rightarrow \mathcal{N}\left(0, 2N_d\left[(V \otimes V) + V \otimes M/\alpha + M \otimes V/\alpha\right]\right)$$

which implies the result. If $\alpha = 0$, then

$$\operatorname{var}\operatorname{vec}\left(n^{-1/2}\textstyle\sum_i (X_i-\mu_i)(X_i-\mu_i)'\right) \to 0$$

so that $\operatorname{vec}\left(n^{-1/2}\sum_i (X_i-\mu_i)(X_i-\mu_i)' - \nu_n V\right) = o_p(1)$. Therefore, we have:

$$\begin{aligned}\sqrt{n}\operatorname{vec}(P/n - \mathbb{E}[P/n]) &= n^{-1/2}\operatorname{vec}\left(\textstyle\sum_i (X_i-\mu_i)(X_i-\mu_i)' - \nu_n V + \sum_i (X_i\mu_i' + \mu_i X_i' - 2\mu_i\mu_i')\right)\\ &= n^{-1/2}\textstyle\sum_i \operatorname{vec}\left(X_i\mu_i' + \mu_i X_i' - 2\mu_i\mu_i'\right) + o_p(1)\\ &= n^{-1/2}\textstyle\sum_i \left((X_i-\mu_i)\otimes\mu_i + \mu_i\otimes(X_i-\mu_i)\right) + o_p(1)\end{aligned} \tag{C.4}$$

Now,

$$\begin{aligned}&\mathbb{E}[(X_i-\mu_i)\otimes\mu_i + \mu_i\otimes(X_i-\mu_i)]^2\\ &\quad= V\otimes\mu_i\mu_i' + \mathbb{E}[(X_i-\mu_i)\mu_i'\otimes\mu_i(X_i-\mu_i)'] + [\mu_i(X_i-\mu_i)'\otimes(X_i-\mu_i)\mu_i'] + \mu_i\mu_i'\otimes V\\ &\quad= (I+K_{dd})\left(V\otimes\mu_i\mu_i' + \mu_i\mu_i'\otimes V\right)\end{aligned}$$

where the last line uses the identity $ab'\otimes ba' = a\otimes(bb'\otimes a) = K_{dd}(bb'\otimes aa')$ for any vectors $a, b\in\mathbb{R}^d$ that follows from Equations (C.2) and (C.3a). Hence

$$\sum_i \left((X_i-\mu_i)\otimes\mu_i + \mu_i\otimes(X_i-\mu_i)\right) \sim \mathcal{N}_{d^2}\left(0, (I+K_{dd})\left(V\otimes M_n + M_n\otimes V\right)\right)$$

which, combined with (C.4), yields the result. □

Corollary C.1. *Consider the model* (3.1)–(3.2) *and suppose Assumptions ER and MI hold. Then:*

$$\begin{aligned}\sqrt{n}\left(\operatorname{vec}(S) - \Omega\right) &\Rightarrow \mathcal{N}_4\left(0, \frac{1}{1-\alpha_k-\alpha_\ell}2N_2(\Omega\otimes\Omega)\right)\\ \sqrt{n}\left(\operatorname{vec}(T) - \alpha_k\Omega - \frac{\lambda_n}{a'\Omega^{-1}a}aa'\right) &\Rightarrow \mathcal{N}_4\left(2N_2\Phi\right)\end{aligned}$$

where $\Phi = \alpha_k\Omega\otimes\Omega + \frac{\lambda}{a'\Omega^{-1}a}\Omega\otimes(aa') + \frac{\lambda}{a'\Omega^{-1}a}(aa')\otimes\Omega$, *and* N_2 *is the symmetrizer matrix.*

Proof. The result follows from Lemma C.1 (ii). □

Lemma C.2. *Consider an invertible matrix* $V\in\mathbb{R}^{d\times d}$, *a vector* $m\in\mathbb{R}^d$ *and a constant* c. *Then:*

$$\begin{aligned}\left(V\otimes V + c(mm')\otimes(mm')\right)^{-1} &= V^{-1}\otimes V^{-1} - \frac{c(V^{-1}mm'V^{-1})\otimes(V^{-1}mm'V^{-1})}{1+c(m'V^{-1}m)^2}\\ (D_d'(V\otimes V + c(mm')\otimes(mm'))D_d)^{-1} &= L_dN_d\left(V\otimes V + c(mm')\otimes(mm')\right)^{-1}N_dL_d'\\ (L_dN_d'(V\otimes V + c(mm')\otimes(mm'))L_d')^{-1} &= D_d'\left(V\otimes V + c(mm')\otimes(mm')\right)^{-1}D_d\end{aligned}$$

Proof. The first identity can be checked by direct calculation. The second identity follows from Lemma 4.4 in Magnus and Neudecker (1980). □

Lemma C.3. *Consider the quadratic form $Q = (V + M)'P(M + V)$, where $P \in \mathbb{R}^{n\times n}$ is symmetric with $\operatorname{tr}(P^2) = r$, $V, M \in \mathbb{R}^{n\times g}$, the rows $v_i \sim [0, \Omega]$ of V are iid with finite fourth moments, and M is non-random.*

(i) *The variance of Q is given by:*

$$\begin{aligned}\operatorname{var}(\operatorname{vec}(Q)) = (I_g + K_{gg})(M'PM \otimes \Omega + \Omega \otimes M'PM + (r - d'd)\Omega \otimes \Omega) \\ + d'd\left[\mathbb{E}(vv') \otimes (vv') - \operatorname{vec}(\Omega)\operatorname{vec}(\Omega)'\right] \\ + \mathbb{E}[vv' \otimes (\overline{m}v' + v\overline{m}')] + \mathbb{E}[(\overline{m}v' + v\overline{m}') \otimes vv']\end{aligned}$$

where $\overline{m} = M'P\operatorname{diag}(P)$ and $d = \operatorname{diag}(P)$.

(ii) *Suppose in addition that for some constant D,*

(a) $\sup_{i\geq 1}\|m_i\| < D < \infty$;

(b) $M'PM/n \xrightarrow{p} \Lambda$;

(c) $\overline{m}/n \to \mu$;

(d) $r/n \to \tau_r$ *and* $\operatorname{tr}(P) = \tau_P + o(n^{-1/2})$;

(e) $d'd/n \to \delta$;

(f) $\sup_n \sup_{1\leq i\leq n} \sum_{s=1}^n |p_{si}| < D < \infty$.

Then:

$$n^{-1/2}\operatorname{vec}(Q - M'PM - \operatorname{tr}(P)\Omega) \Rightarrow \mathcal{N}(0, \operatorname{plim}(\operatorname{var}(\operatorname{vec}(Q))/n))$$

Proof. Proof of part (i) follows by a tedious but straightforward calculation, and it is available at `http://www.people.fas.harvard.edu/~mkolesar/papers/re-supplemental.pdf`. Proof of part (ii) follows from part (i) and Theorem 1 in van Hasselt (2010). □

C.3 Likelihood derivations

***Derivation of Equations* (3.8), (3.9), *and* (3.13).** It follows from equations (3.6) and (3.7) that the densities of the statistics $\hat{\Pi}_1$ and S are, up to a constant, given by:

$$\begin{aligned} f_{\hat{\Pi}_1}(\hat{\Pi}_1;\beta,\eta_n,\Omega) &= |\Omega|^{-k_n/2}\exp\left(-\frac{n}{2}\left(\operatorname{tr}(\Omega^{-1}T)+\eta_n'\eta_n-2(a'\Omega^{-1}a)^{-1/2}\eta_n'\hat{\Pi}_1\Omega^{-1}a\right)\right) \\ f_S(S;\Omega) &= |\Omega|^{-(n-k_n-\ell_n)/2}|S|^{(n-k_n-\ell_n-3)/2}e^{-\frac{n-k_n-\ell_n}{2}\operatorname{tr}(\Omega^{-1}S)} \end{aligned} \tag{C.5}$$

Multiplying these densities yields the limited information likelihood in Equation (3.8). Its score is given by:

$$\begin{aligned} \mathcal{S}_\beta(\beta,\eta_n,\Omega) &= n\frac{\eta_n'\hat{\Pi}_1\Omega^{-1}e_1}{(a\Omega^{-1}a)^{1/2}}-n\frac{\eta_n'\hat{\Pi}_1\Omega^{-1}a}{(a\Omega^{-1}a)^{3/2}}a'\Omega^{-1}e_1 \\ \mathcal{S}_{\eta_n}(\beta,\eta_n,\Omega) &= n\left(\frac{\hat{\Pi}_1\Omega^{-1}a}{(a'\Omega^{-1}a)^{1/2}}-\eta_n\right) \\ \mathcal{S}_{\operatorname{vech}(\Omega)}(\beta,\eta_n,\Omega) &= \frac{1}{2}\tilde{D}'\left[\operatorname{vec}\left(Y_\perp'Y_\perp-(n-\ell_n)\Omega\right)-\frac{2n}{(a'\Omega^{-1}a)^{1/2}}\hat{\Pi}_1'\eta_n\otimes a+\frac{n\eta_n'\hat{\Pi}_1\Omega^{-1}a}{(a'\Omega^{-1}a)^{3/2}}a\otimes a\right] \end{aligned}$$

Taking a second derivative and expectations then after some algebra yields the formula for the information matrix in Equation (3.9).

The second score equation implies that $\hat{\eta}(\Omega,\beta)=\frac{\hat{\Pi}_1\Omega^{-1}a}{(a'\Omega^{-1}a)^{1/2}}$. Therefore, the likelihood with η_n concentrated out is given by:

$$\mathcal{L}_{\text{LI},n}(\beta,\hat{\eta}(\beta,\Omega),\Omega)\propto|\Omega|^{-(n-\ell_n)/2}e^{-\frac{1}{2}\left((n-k_n-\ell_n)\operatorname{tr}(\Omega^{-1}S)+nQ_{\mathcal{S}}(\beta,\Omega)\right)}$$

where I use the identity (C.1a). I concentrate out Ω next. Taking a derivative with respect to Ω, setting it to zero and pre- and post-multiplying the expression by $\hat{\Omega}(\beta)$ yields: yields:

$$0=(n-k_n-\ell_n)S-(n-\ell_n)\hat{\Omega}(\beta)+\frac{nQ_{\mathcal{S}}(\beta,\hat{\Omega}(\beta))}{b'\hat{\Omega}(\beta)b}\hat{\Omega}(\beta)bb'\hat{\Omega}(\beta) \tag{C.6}$$

Post-multiplying this expression by b, and pre- and post-multiplying it by b' and b yields:

$$\hat{\Omega}(\beta)b=\frac{n-k_n-\ell_n}{n-\ell_n-nQ_{\mathcal{S}}(\beta,\hat{\Omega}(\beta))}Sb \qquad\qquad b'\hat{\Omega}(\beta)b=b'\frac{Y_\perp'Y_\perp}{n-\ell_n}b$$

It follows that $Q_{\mathcal{S}}(\beta,\hat{\Omega}(\beta))=(n-\ell_n)Q_{\mathcal{S}}(\beta,S)/(nQ_{\mathcal{S}}(\beta,S)+n-k_n-\ell_n)$. Plugging these expressions back into Equation (C.6) and simplifying yields:

$$\begin{aligned} \hat{\Omega}(\beta) &= \frac{n-k_n-\ell_n}{n-\ell_n}S+\frac{nQ_{\mathcal{S}}(\beta,S)(nQ_{\mathcal{S}}(\beta,S)+n-k_n-\ell_n)}{(n-\ell_n)b'Y_\perp'Y_\perp b}Sbb'S \\ &= \frac{n-k_n-\ell_n}{n-\ell_n}S+\frac{nQ_{\mathcal{S}}(\beta,S)}{(n-\ell_n)b'Sb}Sbb'S \end{aligned}$$

where the second line uses the identity $b'Y_\perp'Y_\perp b=(nQ_{\mathcal{S}}(\beta,S)+n-k_n-\ell_n)b'Sb$. It follows that:

$$\begin{aligned} |\hat{\Omega}(\beta)| &= \left(n+(n-k_n-\ell_n)/Q_{\mathcal{S}}(\beta,S)\right)^{-1} \\ (n-k_n-\ell_n)\operatorname{tr}\left(\hat{\Omega}(\beta)^{-1}S\right) &= (n-\ell_n)\left(2-\frac{Q_{\mathcal{S}}(\beta,S)}{(n-k_n-\ell_n)/n+Q_{\mathcal{S}}(\beta,S)}\right) \end{aligned}$$

Therefore, the likelihood with both Ω and η_n concentrated out is given by:

$$\mathcal{L}_{\text{LI},n}(\beta, \hat{\eta}(\beta), \hat{\Omega}(\beta)) \propto (n + (n - k_n - \ell_n)/Q_{\mathcal{S}}(\beta, S))^{(n-\ell_n)/2} e^{-(n-\ell_n)}$$

Maximizing this concentrated likelihood is equivalent to minimizing $Q_{\mathcal{S}}(\beta, S)$, so that the maximum limited information likelihood estimator of β is given by $\hat{\beta}_{\text{LIML}}$. Consequently:

$$\hat{\Omega}_{\text{LIML}} = \frac{n - k_n - \ell_n}{n - \ell_n} S + \frac{n}{n - \ell_n} \frac{m_{\min}}{b'Sb} Sbb'S \tag{C.7}$$

Using the identities (C.1) yields the expression for $\hat{\Omega}_{\text{LIML}}$ in Equation (3.13). To derive $\hat{\eta}_{\text{LIML}}$, note that Equation (C.7) and the Woodbury formula yield $\hat{\Omega}^{-1}_{\text{LIML}} \hat{a}_{\text{LIML}} = (n - \ell_n) S^{-1} \hat{a}_{\text{LIML}} / (n - k_n - \ell_n)$. Consequently:

$$\hat{\eta}_{\text{LIML}} = \sqrt{\frac{n - \ell_n}{n - k_n - \ell_n}} \frac{\hat{\Pi}_1 S^{-1} \hat{a}_{\text{LIML}}}{\sqrt{\hat{a}'_{\text{LIML}} S^{-1} \hat{a}_{\text{LIML}}}}$$

The expression for $\hat{\lambda}_{\text{LIML}}$ in Equation (3.13) then follows.

It remains to derive the probability limits in Equation (3.13). By continuity of the trace operator, and Corollary C.1:

$$\text{tr}(S^{-1}T) \xrightarrow{p} 2\alpha_k + \lambda$$

Secondly, again by Corollary C.1:

$$m_{\min} = \frac{\hat{b}'_{\text{LIML}} T \hat{b}_{\text{LIML}}}{\hat{b}'_{\text{LIML}} S \hat{b}_{\text{LIML}}} \xrightarrow{p} \frac{\alpha_k b'\Omega b}{b'\Omega b} = \alpha_k$$

Hence,

$$m_{\max} \xrightarrow{p} \lambda + \alpha_k \tag{C.8}$$

The probability limit for $\hat{\lambda}_{\text{LIML}}$ follows. The probability limit for Ω follows from Lemma (ii), Equation (C.8), and the Slutsky's Theorem. □

Derivation of Equation (3.17)**.** The distribution of the $\hat{\Pi}_1$ with η_n integrated out according to the random-effects prior (3.16) is given by:

$$\text{vec}(\hat{\Pi}_1) \sim \mathcal{N}_{2k_n}\left(0, \frac{1}{n}\left(\frac{n}{k_n}\frac{\lambda}{a'\Omega^{-1}a} aa' + \Omega\right) \otimes I_{k_n}\right)$$

It follows from the formulae for multivariate Normal density and the identities

$$\text{tr}\left(\left(\Omega + \frac{n}{k_n}\frac{\lambda}{a'\Omega^{-1}a} aa'\right)^{-1} T\right) = \text{tr}(\Omega^{-1}T) - \frac{\lambda}{\frac{k_n}{n} + \lambda} Q_{\mathcal{T}}(\beta, \Omega)$$
$$\left|\Omega + \frac{n}{k_n}\frac{\lambda}{a'\Omega^{-1}a} aa'\right| = |\Omega|\left(1 + \frac{n}{k_n}\lambda\right)$$

that

$$f_{\hat{\Gamma}\mathrm{I}_1}(\hat{\Gamma}\mathrm{I}_1;\beta,\lambda,\Omega) \propto |\Omega|^{-k_n/2}\left(1+\frac{n}{k_n}\lambda\right)^{-k_n/2} e^{-\frac{n}{2}\operatorname{tr}(\Omega^{-1}T)+\frac{n}{2}\frac{\lambda}{\frac{n}{k_n}+\lambda}Q_T(\beta,\Omega)} \tag{C.9}$$

Combining this expression with the density for S given by Equation (C.5) and the fact that $\hat{\Gamma}\mathrm{I}_1^*$ and S are independent then yields the result. □

Derivation of Equation (3.25)**.** The distribution of $\hat{\Gamma}\mathrm{I}_1$ with the means integrated out using the random effects prior (3.24) is given by:

$$\hat{\Gamma}\mathrm{I}_1^* \sim \mathcal{N}_{2k_n}\left(0, \left(\frac{n}{k_n}\Xi(\beta,\Lambda_{11},\Lambda_{22})+\Omega\right)\otimes I_{k_n}\right) \qquad \Xi(\beta,\Lambda_{11},\Lambda_{22}) = \Gamma\Lambda\Gamma' = \begin{pmatrix}\Lambda_{11}+\Lambda_{22}\beta^2 & \Lambda_{22}\beta \\ \Lambda_{22}\beta & \Lambda_{22}\end{pmatrix}$$

Define $\tilde{\Lambda} = \frac{n}{k_n}\Lambda$, and let $V = \Omega + (\tilde{\Lambda}_{22}aa' + \tilde{\Lambda}_{11}e_1e_1')$. Then, using the formula for a multivariate Normal random variable, the density of $\hat{\Gamma}\mathrm{I}_1^*$ is given by:

$$f_{\hat{\Gamma}\mathrm{I}^*}(\hat{\Gamma}\mathrm{I}_1^*;\beta,\Lambda,\Omega) = (2\mathrm{pi})^{-k_n}|V|^{-k_n/2}e^{-\frac{n}{2}\operatorname{tr}(V^{-1}T)} \tag{C.10}$$

Applying the Woodbury identity twice yields:

$$\begin{aligned}|V| &= |\Omega|\left(1+\tilde{\Lambda}_{22}a'\Omega^{-1}a+\tilde{\Lambda}_{11}e_1'\Omega^{-1}e_1+\tilde{\Lambda}_{11}\tilde{\Lambda}_{22}\left(a'\Omega^{-1}ae_1'\Omega^{-1}e_1-(e_1'\Omega^{-1}a)^2\right)\right)\\ &= |\Omega|+\tilde{\Lambda}_{22}b'\Omega b+\tilde{\Lambda}_{11}\Omega_{22}+\tilde{\Lambda}_{11}\tilde{\Lambda}_{22} = \frac{n^2}{k_n^2}D(\beta,\Lambda_{11},\Lambda_{22},\Omega)\end{aligned} \tag{C.11}$$

where I use (C.1b) and $a'\Omega^{-1}ae_1'\Omega^{-1}e_1 - (e_1'\Omega^{-1}a)^2 = |\Gamma'\Omega^{-1}\Gamma| = |\Omega|^{-1}$ in the second line. Secondly, applying Woodbury identity after some algebra yields:

$$\operatorname{tr}\left(V^{-1}T\right) = \operatorname{tr}(\Omega^{-1}T) - \frac{\tilde{\Lambda}_{22}a'Za+Z_{11}\tilde{\Lambda}_{11}+\tilde{\Lambda}_{22}\tilde{\Lambda}_{11}\frac{1}{|\Omega|}\left(Z_{11}\Omega_{11}+2Z_{12}\Omega_{12}+\Omega_{22}Z_{22}\right)}{1+\frac{1}{|\Omega|}\left(\tilde{\Lambda}_{22}b'\Omega b+\tilde{\Lambda}_{11}\Omega_{22}+\tilde{\Lambda}_{11}\tilde{\Lambda}_{22}\right)}$$

where $Z = \Omega^{-1}T\Omega^{-1}$. Applying (C.1e) yields:

$$\operatorname{tr}\left(V^{-1}T\right) = \frac{k_n^2/n^2}{D(\beta,\Lambda_{11},\Lambda_{22},\Omega)}\left(|\Omega|\operatorname{tr}(\Omega^{-1}T)+\tilde{\Lambda}_{22}b'Tb+\tilde{\Lambda}_{11}T_{22}\right) \tag{C.12}$$

Plugging (C.11) and (C.12) into (C.10) yields:

$$f_{\hat{\Gamma}\mathrm{I}_1^*}(\hat{\Gamma}\mathrm{I}_1^*;\beta,\Lambda,\Omega) = (2\mathrm{pi})^{-k_n}D(\beta,\Lambda_{11},\Lambda_{22},\Omega)^{-k_n/2}e^{-\frac{k_n(k_n/n)}{2D(\beta,\Lambda_{11},\Lambda_{22},\Omega)}\left(|\Omega|\operatorname{tr}(\Omega^{-1}T)+\tilde{\Lambda}_{22}b'Tb+\tilde{\Lambda}_{11}T_{22}\right)}$$

Combining this density with the density for S given in Equation (C.5) then yields the result. □

C.4 Proofs

Proof of Proposition 3.1. The density of T is proportional to (Moreira, 2009, Theorem 4.1):

$$f_T(T \mid \beta,\lambda_n,\Omega) = e^{-\frac{n}{2}\left(\lambda_n+\operatorname{tr}(\Omega^{-1}T)\right)}|\Omega|^{-k_n/2}\left(n\sqrt{\lambda_n Q_T(\beta)}\right)^{-(k_n-2)/2} I_{(k_n-2)/2}\left(n\sqrt{\lambda_n Q_T(\beta)}\right) \tag{C.13}$$

where $I_\nu(\cdot)$ is modified Bessel function of the first kind of order ν. Using the integral representation of the Bessel function (Abramowitz and Stegun, 1965),

$$I_\nu(t) = \frac{(t/2)^\nu}{\pi^{1/2}\bar{\Gamma}(\nu+1/2)} G_{2\nu+2}(t) \qquad\qquad G_\nu(t) = \int_{[-1,1]} e^{ts}(1-s^2)^{(\nu-3)/2}\,\mathrm{d}s$$

where $\bar{\Gamma}$ is the gamma function. The density (C.13) can be written (up to a constant) as:

$$f_T(T \mid \beta, \lambda_n, \Omega) = e^{-\frac{n}{2}\left(\lambda_n + \operatorname{tr}(\Omega^{-1}T)\right)} |\Omega|^{-k_n/2}\, G_{k_n}\left(n\sqrt{\lambda_n Q_{\mathcal{T}}(\beta)}\right)$$

Combining this expression with the density for S, given in Equation (C.5), yields the invariant likelihood:

$$\log \mathcal{L}_{\text{INV},n}(\beta, \lambda_n, \Omega; S, T) \propto -\frac{1}{2}\left((n-\ell_n)\log|\Omega| + \operatorname{tr}(\Omega^{-1}Y_\perp' Y_\perp) + n\lambda_n - \log G_{k_n}(n\sqrt{\lambda_n Q_{\mathcal{T}}(\beta,\Omega)})\right)$$

Fix λ_n. Dropping the k_n index from the G function to avoid clutter, the derivative with respect to Ω is given by:

$$\frac{\partial \log \mathcal{L}_{\text{INV},n}}{\partial \Omega} = \frac{1}{2}\left[\Omega^{-1}Y_\perp' Y_\perp \Omega^{-1} - (n-\ell_n)\Omega^{-1} - \frac{G'(\cdot)}{2G(\cdot)} \frac{n\lambda_n^{1/2}}{Q_{\mathcal{T}}(\beta,\Omega)^{1/2}}\left(\Omega^{-1}T\Omega^{-1} - \frac{Q_{\mathcal{S}}(\beta,\Omega)}{b'\Omega b} bb'\right)\right] \tag{C.14}$$

where the derivative $\partial Q_{\mathcal{T}}(\beta,\Omega)/\partial\Omega$, given by the expression in parentheses, is computed using the identity (C.1a). Denote the ML estimates of β and Ω given λ_n by $(\hat{\beta}_{\lambda_n}, \hat{\Omega}_{\lambda_n})$. Since $G(\cdot)$ is a monotone function, it follows from the expression for the invariant likelihood that:

$$\hat{\beta}_{\lambda_n} = \operatorname*{argmax}_\beta Q_{\mathcal{T}}(\beta, \hat{\Omega}_{\lambda_n}) = \operatorname*{argmin}_\beta Q_{\mathcal{S}}(\beta, \hat{\Omega}_{\lambda_n})$$

Secondly, the derivative (C.14) evaluated at $(\hat{\beta}_{\lambda_n}, \hat{\Omega}_{\lambda_n})$ has to be equal to zero. Pre-multiplying and post-multiplying Equation (C.14) by $\hat{b}'_{\lambda_n}\hat{\Omega}_{\lambda_n}$ and post-multiplying it by $\hat{\Omega}_{\lambda_n}\hat{b}_{\lambda_n}$ therefore yields:

$$(n-\ell_n)\hat{b}'_{\lambda_n}\hat{\Omega}_{\lambda_n}\hat{b}_{\lambda_n} = \hat{b}'_{\lambda_n} Y_\perp' Y_\perp \hat{b}_{\lambda_n} \tag{C.15}$$

This implies:

$$\hat{\beta}_{\lambda_n} = \operatorname*{argmin}_\beta Q_{\mathcal{S}}(\beta, \hat{\Omega}_{\lambda_n}) = \operatorname*{argmin}_\beta Q_{\mathcal{S}}(\beta, Y_\perp' Y_\perp) = \hat{\beta}_{\text{LIML}}$$

as required. □

Proof of Proposition 3.2. It follows from Equation (3.17) that the log-likelihood, parametrized in terms of (ψ, λ, Ω) where $\psi = \Omega^{-1}a$, can be written as:

$$\log \mathcal{L}_{\text{RE},n}(\psi, \lambda, \Omega) = -\frac{1}{2}\left((n-\ell_n)\log|\Omega| + k_n \log\left(1 + \frac{n}{k_n}\lambda\right) + \operatorname{tr}(\Omega^{-1}Y_\perp' Y_\perp) - \frac{n\lambda}{k_n/n+\lambda} Q_{\mathcal{S}}(\psi,\Omega)\right)$$

where $Q_{\mathcal{S}}(\psi, \Omega) = \psi' T \psi / (\psi' \Omega \psi)$. The derivative with respect to λ is given by

$$\frac{\partial}{\partial \lambda} \log \mathcal{L}_{\text{RE},n}(\psi, \lambda, \Omega) = -\frac{1}{2} \frac{k_n}{k_n/n + \lambda} \left(1 - \frac{Q_{\mathcal{S}}(\psi, \Omega)}{k_n/n + \lambda}\right)$$

Now suppose that

$$Q_{\mathcal{S}}(\psi, \Omega) > k_n/n \tag{C.16}$$

Then the ML estimator of λ with ψ and Ω given is given by:

$$\hat{\lambda}_{\psi,\Omega} = Q_{\mathcal{S}}(\psi, \Omega) - k_n/n$$

Therefore, the likelihood with λ concentrated out is given by:

$$\log \mathcal{L}_{\text{RE},n}(\psi, \hat{\lambda}_{\psi,\Omega}, \Omega) \propto -\frac{1}{2} \left((n - \ell_n) \log|\Omega| + k_n \log \left(Q_{\mathcal{S}}(\psi, \Omega)\right) + \text{tr}(\Omega^{-1} Y_\perp' Y_\perp) - n Q_{\mathcal{S}}(\psi, \Omega) \right)$$

The derivative with respect to Ω is given by:

$$\frac{\partial}{\partial \Omega} \log \mathcal{L}_{\text{RE},n}(\psi, \hat{\lambda}_{\psi,\Omega}, \Omega) = \Omega^{-1} Y_\perp' Y_\perp \Omega^{-1} - (n - \ell_n)\Omega^{-1} + \frac{k_n - n Q_{\mathcal{S}}(\psi, \Omega)}{\psi' \Omega \psi} \psi \psi' \tag{C.17}$$

Setting the derivative to zero, and pre-multiplying it by $\hat{\Omega}_\psi$ and $\psi' \hat{\Omega}_\psi$, and post-multiplying it by $\hat{\Omega}_\psi \psi$ yields:

$$\psi' S \psi = \psi' \hat{\Omega}_\psi \psi \qquad\qquad \frac{1}{(n - k_n - \ell_n + n Q_{\mathcal{S}}(\psi, S))} Y_\perp' Y_\perp \psi = \Omega \psi \tag{C.18}$$

where $Q_{\mathcal{S}}(\psi, S) = Q_{\mathcal{S}}(\psi, \hat{\Omega}_\psi)$. Plugging these expressions back into (C.17) yields:

$$(n - \ell_n)\hat{\Omega}_\psi = Y_\perp' Y_\perp + \frac{k_n - n Q_{\mathcal{S}}(\psi, S)}{\psi' S \psi} \frac{1}{(n - k_n - \ell_n + n Q_{\mathcal{S}}(\psi, S))^2} Y_\perp' Y_\perp \psi \psi' Y_\perp' Y_\perp \tag{C.19}$$

Hence:

$$|\hat{\Omega}_\psi| = \frac{1}{(n - \ell_n)} \frac{|Y_\perp' Y_\perp|}{n - k_n - \ell_n + n Q_{\mathcal{S}}(\psi, S)} \qquad \text{tr}(\hat{\Omega}_\psi^{-1} Y_\perp' Y_\perp) = 2(n - \ell_n) - k_n + n Q_{\mathcal{S}}(\psi, S)$$

Therefore, the likelihood with both λ and Ω concentrated out is given by:

$$\log \mathcal{L}_{\text{RE},n}(\psi, \hat{\lambda}_\psi, \hat{\Omega}_\psi) \propto \frac{1}{2} \left((n - \ell_n) \log(n - k_n - \ell_n + n Q_{\mathcal{S}}(\psi, S)) - k_n \log \left(Q_{\mathcal{S}}(\psi, S)\right) \right)$$

This expression is increasing in $Q_{\mathcal{S}}$ if $Q_{\mathcal{S}} > k_n/n$. The maximum is obtained at $Q_{\mathcal{S}}(\hat{\psi}_{\text{RE}}, S) = m_{\max}$, Equation (C.16) holds, and $\hat{\lambda}_{\text{RE}} = m_{\max} - k_n/n$.

The estimator $\hat{\psi}_{\text{RE}}$ is given by the eigenvector that corresponds to the $m_{\max}$, the larger eigenvalue of $S^{-1}T$. Therefore, $S^{-1} T \hat{\psi}_{\text{RE}} = m_{\max} \hat{\psi}_{\text{RE}}$. Secondly, since $Q_{\mathcal{S}}(\hat{\psi}_{\text{RE}}, S) = Q_{\mathcal{T}}(\hat{\beta}_{\text{LIML}}, S)$, it follows that $\hat{\psi}_{\text{RE}} = S^{-1} \hat{a}_{\text{LIML}}$. Combining these two observations yields $Y_\perp' Y_\perp \hat{\psi}_{\text{RE}} = S(n - k_n - \ell_n + n m_{\max})\psi = (n - k_n - \ell_n + n m_{\max}) \hat{a}_{\text{LIML}}$,

and $\hat{\psi}'_{\text{RE}} S \hat{\psi}_{\text{RE}} = \hat{a}'_{\text{LIML}} S \hat{a}_{\text{LIML}}$. Plugging these result into Equations (C.18) and (C.19) yields:

$$\hat{a}_{\text{RE}} = \hat{\Omega}_{\text{RE}} \hat{\psi}_{\text{RE}} = \hat{a}_{\text{LIML}} \qquad\qquad \hat{\Omega}_{\text{RE}} = \frac{1}{n - \ell_n} \left(Y'_\perp Y_\perp - \frac{\hat{\lambda}_{\text{RE}}}{\hat{a}'_{\text{RE}} S^{-1} \hat{a}_{\text{RE}}} \hat{a}_{\text{RE}} \hat{a}'_{\text{RE}} \right)$$

Next, the consistency of $\hat{\lambda}_{\text{RE}}$ follows from Equation (C.8). The consistency of $\hat{\Omega}_{\text{RE}}$ follows by consistency of $\hat{\lambda}_{\text{RE}}$ and $\hat{\beta}_{\text{RE}}$, Corollary C.1, and Slutsky's Theorem. □

Proof of Proposition 3.3. To avoid clutter, I write $(\hat{\beta}, \hat{\lambda}, \hat{\Omega})$ in place of place of $(\hat{\beta}_{\text{RE}}, \hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$ and $\hat{Q}_{\mathcal{S}}$ in $Q_{\mathcal{S}}(\hat{\beta}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$. The score equations based on the random-effects likelihood (3.17) are given by:

$$\mathcal{S}_\beta(\beta, \lambda, \Omega) = \frac{n\lambda}{k_n/n + \lambda} \frac{e'_2 \left(T - Q_{\mathcal{S}}(\beta, \Omega)\Omega\right) b}{b' \Omega b} \tag{C.20a}$$

$$\mathcal{S}_\lambda(\beta, \lambda, \Omega) = -\frac{1}{2} \frac{k_n}{k_n/n + \lambda} \left(1 - \frac{Q_{\mathcal{T}}(\beta, \Omega)}{k_n/n + \lambda}\right) \tag{C.20b}$$

$$\mathcal{S}_\Omega(\beta, \lambda, \Omega) = \frac{1}{2} D' \operatorname{vec} \left[\Omega^{-1} Y'_\perp Y_\perp \Omega^{-1} - (n - \ell_n)\Omega^{-1} - \frac{n\lambda}{k_n/n + \lambda} \left(\Omega^{-1} T \Omega^{-1} - \frac{Q_{\mathcal{S}}}{b'\Omega b} bb' \right) \right] \tag{C.20c}$$

The Hessian, evaluated at ML estimates, is given by:

$$H_{\text{RE}}(\hat{\beta}, \hat{\lambda}, \hat{\Omega}) = \begin{pmatrix} \frac{n\hat{\lambda}}{(k_n/n + \hat{\lambda}) \hat{b}' \hat{\Omega} \hat{b}} (\hat{Q}_{\mathcal{S}} \hat{\Omega}_{22} - T_{22}) & 0 & \hat{H}_{1,3:5} \\ 0 & -\frac{1}{2} \frac{k_n}{(k_n/n + \hat{\lambda})^2} & \hat{H}_{2,3:5} \\ \hat{H}'_{1,3:5} & \hat{H}'_{2,3:5} & \hat{H}_{3:6,3:5} \end{pmatrix}$$

where

$$\begin{aligned}
H_{1,3:5} &= -\frac{1}{2} \frac{\hat{c}(n - \ell_n)}{\hat{b}' \hat{\Omega} \hat{b}} \left(2 \frac{e'_2 \hat{\Omega} \hat{b}}{\hat{b}' \hat{\Omega} \hat{b}} \hat{b} \otimes \hat{b} - \hat{b} \otimes e_2 - e_2 \otimes b \right)' D \\
\hat{H}_{2,3:5} &= -\frac{1}{2} \frac{k_n}{(k_n/n + \hat{\lambda})^2} \left(\frac{\hat{Q}_{\mathcal{S}}}{\hat{b} \hat{\Omega} \hat{b}} \hat{b} \otimes \hat{b} - \operatorname{vec}(\hat{\Omega}^{-1} T \hat{\Omega}^{-1}) \right)' D \\
\hat{H}_{3:5,3:5} &= -\frac{(n - \ell_n)}{2} D' \left(\left(\hat{\Omega}^{-1} - \frac{\hat{c} \hat{b} \hat{b}'}{\hat{b}' \hat{\Omega} \hat{b}} \right) \otimes \left(\hat{\Omega}^{-1} - \frac{\hat{c} \hat{b} \hat{b}'}{\hat{b}' \hat{\Omega} \hat{b}} \right) - (2\hat{c} - \hat{c}^2) \frac{\hat{b} \hat{b}'}{\hat{b}' \hat{\Omega} \hat{b}} \otimes \frac{\hat{b} \hat{b}'}{\hat{b}' \hat{\Omega} \hat{b}} \right) D
\end{aligned}$$

By the formula for block inverses, the upper 2×2 submatrix of the inverse Hessian is given by:

$$H^{1:2,1:2}(\hat{\beta}, \hat{\lambda}, \hat{\Omega}) = \left(\hat{H}_{1:2,1:2} - \hat{H}_{1:2,3:5} \hat{H}^{-1}_{3:5,3:5} \hat{H}'_{1:2,3:5} \right)^{-1} \tag{C.21}$$

Applying Lemma C.2 yields:

$$\hat{H}^{-1}_{3:5,3:5} = -\frac{2}{n - \ell_n} LN \left[\left(\hat{\Omega} + \frac{\hat{c}}{1 - \hat{c}} \frac{\hat{\Omega} \hat{b} \hat{b}' \hat{\Omega}}{\hat{b}' \hat{\Omega} \hat{b}} \right) \otimes \left(\hat{\Omega} + \frac{\hat{c}}{1 - \hat{c}} \frac{\hat{\Omega} \hat{b} \hat{b}' \hat{\Omega}}{\hat{b}' \hat{\Omega} \hat{b}} \right) + \frac{\hat{c}^2 - 2\hat{c}}{(1 - \hat{c})^2} \frac{\hat{\Omega} \hat{b} \hat{b}' \hat{\Omega} \otimes \hat{\Omega} \hat{b} \hat{b}' \hat{\Omega}}{(\hat{b}' \hat{\Omega} \hat{b})^2} \right] NL'$$

where L is the elimination matrix and N is the symmetrizer matrix. It follows that:

$$\hat{H}_{1:2,3:5} \hat{H}^{-1}_{3:5,3:5} \hat{H}'_{1:2,3:5} = -\frac{(n - \ell_n)\hat{c}^2}{1 - \hat{c}} \frac{|\Omega|}{(b' \Omega b)^2}$$

Finally, since $\hat{H}_{2,3:6}\hat{H}_{3:6,3:6}^{-1}\hat{H}_{1,3:6}' = 0$, Equation (C.21) combined with the expression above yields:

$$\hat{H}_{\text{RE}}^{11} = \left(\hat{H}_{11} - \hat{H}_{1,3:5}\hat{H}_{3:5,3:5}^{-1}\hat{H}_{1,3:5}'\right)^{-1} = \frac{\hat{b}'\hat{\Omega}\hat{b}(\hat{\lambda} + k_n/n)}{n\hat{\lambda}}\left(\hat{Q}_{\mathcal{S}}\hat{\Omega}_{22} - T_{22} + \frac{\hat{c}}{1-\hat{c}}\frac{\hat{Q}_{\mathcal{S}}}{\hat{a}'\hat{\Omega}^{-1}\hat{a}}\right)^{-1}$$

which yields the result. Now, consider its probability limit. We have:

$$\hat{Q}_{\mathcal{S}} = (n-\ell_n)\frac{\hat{b}'T\hat{b}}{\hat{b}'Y_{\perp}'Y_{\perp}\hat{b}} = \frac{(n-\ell_n)\hat{b}'T\hat{b}}{(n-k_n-\ell_n)\hat{b}'S\hat{b} + n\hat{b}'T\hat{b}} = \left(\frac{1}{1-\ell_n/n} + \frac{n-k_n-\ell_n}{(n-\ell_n)m_{\min}}\right)^{-1} \xrightarrow{p} \alpha_k$$

since $m_{\min} \xrightarrow{p} \alpha_k$. Hence:

$$\frac{\hat{c}}{1-\hat{c}} \xrightarrow{p} \frac{\alpha_k\lambda}{\alpha_k(1-\alpha_\ell) + (1-\alpha_k-\alpha_\ell)\lambda}$$

So that:

$$\begin{aligned}-n\hat{H}_{\text{RE}}^{11} &\xrightarrow{p} -\frac{b'\Omega b(\alpha_K+\lambda)}{\lambda}\left(-\frac{\lambda}{a'\Omega^{-1}a} + \frac{\lambda\alpha_K^2}{a'\Omega^{-1}a\left((1-\alpha_K-\alpha_\ell)\lambda + (1-\alpha_\ell)\alpha_K\right)}\right)^{-1} \\ &= \frac{\Sigma_{11}a'\Omega^{-1}a}{\lambda^2}\left(\lambda + \frac{(1-\alpha_\ell)\alpha_K}{1-\alpha_\ell-\alpha_K}\right) = \mathcal{V}_{\text{LIML},N}\end{aligned}$$

which completes the proof. □

Proof of Proposition 3.4. The objective function evaluates as:

$$\begin{aligned}\mathcal{Q}_n(\beta,\lambda,\Omega;\hat{W}_n) = \frac{n-k_n-\ell_n}{n}&\text{tr}(I_2 - 2S^{-1}\Omega + (S^{-1}\Omega)^2) + (k_n/n)^{-1}\,\text{tr}((TS^{-1})^2 - 2(k_n/n)S^{-1}TS^{-1}\Omega \\ &+ (k_n/n)^2(S^{-1}\Omega)^2) + \frac{n\lambda^2}{k_n}\left(\frac{a'S^{-1}a}{a'\Omega^{-1}a}\right)^2 - \frac{2n\lambda}{k_n}\frac{a'S^{-1}(T-(k_n/n)\Omega)S^{-1}a}{a'\Omega^{-1}a} \quad \text{(C.22)}\end{aligned}$$

Denote the minimum distance estimates based on minimizing this objective function by $(\hat{\beta},\hat{\lambda},\hat{\Omega})$. Let $C = S^{-1}((k_n/n)\Omega - T)S^{-1}$. $(\hat{\beta},\hat{\lambda},\hat{\Omega})$ have to satisfy the first-order conditions:

$$0 = \hat{a}'\left(C - \frac{\hat{a}'C\hat{a}}{\hat{a}'\hat{\Omega}^{-1}\hat{a}}\hat{\Omega}^{-1} + \hat{\lambda}\frac{\hat{a}'S^{-1}\hat{a}}{\hat{a}\hat{\Omega}^{-1}\hat{a}}S^{-1} - \hat{\lambda}\frac{\hat{a}'S^{-1}\hat{a}}{\hat{a}\hat{\Omega}^{-1}\hat{a}}\frac{\hat{a}'S^{-1}\hat{a}}{\hat{a}\hat{\Omega}^{-1}\hat{a}}\hat{\Omega}^{-1}\right)e_1 \quad \text{(C.23)}$$

$$0 = \frac{\hat{a}\hat{\Omega}^{-1}\hat{a}}{\hat{a}'S^{-1}\hat{a}}\frac{\hat{a}'C\hat{a}}{\hat{a}'S^{-1}\hat{a}} + \hat{\lambda} \quad \text{(C.24)}$$

$$\begin{aligned}0 = (1-\ell_n/n)S^{-1}\hat{\Omega}S^{-1} - S^{-1}TS^{-1} - \frac{n-k_n-\ell_n}{n}S^{-1} + \\ + \hat{\lambda}\frac{S^{-1}\hat{a}\hat{a}'S^{-1}}{\hat{a}'\hat{\Omega}^{-1}\hat{a}} + \frac{\hat{\lambda}}{(k_n/n)(\hat{a}'\hat{\Omega}^{-1}\hat{a})^2}\left(\hat{a}'C\hat{a} + \frac{\hat{\lambda}(\hat{a}'S^{-1}\hat{a})^2}{(\hat{a}'\hat{\Omega}^{-1}\hat{a})}\right)\hat{\Omega}^{-1}\hat{a}\hat{a}'\hat{\Omega}^{-1} \quad \text{(C.25)}\end{aligned}$$

Combining (C.24) with (C.23) and with the fact $(\hat{\beta},\hat{\lambda},\hat{\Omega})$ have to minimize the objective function implies:

$$0 = \hat{a}'\left(C - \frac{\hat{a}'C\hat{a}}{\hat{a}'S^{-1}\hat{a}}S^{-1}\right)e_1 \qquad \beta = \underset{\beta}{\text{argmax}}\left(\frac{\hat{a}'C\hat{a}}{\hat{a}'S^{-1}\hat{a}}\right)^2 \quad \text{(C.26)}$$

Combining (C.24) and (C.25) yields:

$$0 = (1-\ell_n/n)S^{-1}\hat{\Omega}S^{-1} - S^{-1}TS^{-1} - (n-k_n-\ell_n)S^{-1}/n - \frac{\hat{a}'C\hat{a}}{(\hat{a}'S^{-1}\hat{a})^2}S^{-1}\hat{a}\hat{a}'S^{-1} \quad \text{(C.27)}$$

Pre-multiplying Equation (C.27) by $\hat{a}'$ and post-multiplying it by $\hat{a}$ yields:

$$\hat{a}S^{-1}\hat{a} = \hat{a}S^{-1}\hat{\Omega}S^{-1}\hat{a}$$

Combining this result with (C.26) yields:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left(Q_{\mathcal{T}}(\beta, S^{-1}) - k_n/n\right)^2 = \underset{\beta}{\operatorname{argmax}}\, Q_{\mathcal{T}}(\beta, S^{-1}) = \hat{\beta}_{\text{RE}}$$

The first expression follows since by the identity (C.1a), $\operatorname{tr}(S^{-1}T) > 2k_n/n$ implies $Q_{\mathcal{T}}(\beta, S) - k_n/n > k_n/nQ_{\mathcal{S}}(\beta, S)$, and $\min_\beta Q_{\mathcal{S}}(\beta, S) = \min_\beta Q_{\mathcal{T}}(\beta, S) = m_{\min}$. Consequently:

$$\frac{\hat{a}'C\hat{a}}{\hat{a}'S^{-1}\hat{a}} = k_n/n - m_{\max} \tag{C.28}$$

Pre- and post-multiplying (C.27) by S^{-1}, and combining it with this result yields:

$$\hat{\Omega} = \frac{1}{n-\ell_n}\left[Y_\perp' Y_\perp - \frac{n(m_{\max} - k_n/n)}{\hat{a}'S^{-1}\hat{a}}\hat{a}\hat{a}'\right] \tag{C.29}$$

Hence:

$$\hat{\Omega}^{-1} = (n-\ell_n)\left(Y_\perp' Y_\perp^{-1} + \frac{n(m_{\max} - k_n/n)Y_\perp' Y_\perp^{-1} a a' Y_\perp' Y_\perp^{-1}}{a'S^{-1}a - n(m_{\max} - k_n/n)aY_\perp' Y_\perp^{-1} a}\right)$$

Therefore, using Equations (C.1):

$$\begin{aligned}
\frac{\hat{a}\hat{\Omega}^{-1}\hat{a}}{\hat{a}'S^{-1}\hat{a}} &= (n-\ell_n)\frac{\hat{a}'Y_\perp' Y_\perp^{-1}\hat{a}}{\hat{a}S^{-1}\hat{a} - n(m_{\max} - k_n/n)\hat{a}Y_\perp' Y_\perp^{-1}\hat{a}} \\
&= (n-\ell_n)\frac{b'Y_\perp' Y_\perp b}{\frac{|Y_\perp' Y_\perp|}{|S|}bSb + n(m_{\max} - k_n/n)bY_\perp' Y_\perp b} \\
&= (n-\ell_n)\left(\frac{|Y_\perp' Y_\perp|}{|S|}\left(\frac{b'Y_\perp' Y_\perp b}{bSb}\right)^{-1} - n(m_{\max} - k_n/n)\right)^{-1} \\
&= (n-\ell_n)\left(n(1+\alpha_\ell+\alpha_k+m_{\max}) - n(m_{\max} - k_n/n)\right)^{-1} = 1
\end{aligned} \tag{C.30}$$

Combining this result with (C.24) and (C.28) yields:

$$\hat{\lambda} = m_{\max} - (k_n/n) = \hat{\lambda}_{\text{RE}}$$

and plugging this into (C.29) then finally yields

$$\hat{\Omega} = \hat{\Omega}_{\text{RE}} \qquad \square$$

Proof of Proposition 3.5. Consider the parametrization $(\beta, \Lambda_{22}, \Omega)$ where $\Lambda_{22} = \lambda/(a'\Omega^{-1}a)$, so that the moment conditions are given by:

$$\mathbb{E}\begin{pmatrix} \operatorname{vech}(S-\Omega) \\ \operatorname{vech}(T - \alpha_k\Omega - aa'\Lambda_{22}) \end{pmatrix} = 0 \tag{C.31}$$

Since the reparametrization is one-to-one, if the weight matrix $\hat{W}_n$ is optimal under this parametrization, it

will also be optimal under the original parametrization. Under this parametrization, by Corollary C.1, the asymptotic variance of the moment conditions is given by:

$$\Delta = \begin{pmatrix} 2L_2N_2\frac{(\Omega\otimes\Omega)}{1-\alpha_k-\alpha_\ell}L_2' & 0 \\ 0 & 2L_2N_2\left[\alpha_k\Omega\otimes\Omega+\Omega\otimes(mm')+(mm')\otimes\Omega\right]L_2' \end{pmatrix}$$

where $m = \Lambda_{22}^{1/2}a$. The derivative of the moment condition (C.31) is given by:

$$G = -\begin{pmatrix} 0 & I_3 \\ L_2(m\otimes M + M\otimes m) & \alpha_k I_3 \end{pmatrix} \qquad M = \frac{\mathrm{d}m}{\mathrm{d}(\beta,\lambda)} = \begin{pmatrix} \Lambda_{22}^{1/2}e_1 & \frac{1}{2\Lambda^{1/2}}a \end{pmatrix}$$

Let W_t, $t = c\Lambda_{22}$ denote the probability limit of $\hat{W}_n$ given in the statement of the Proposition. This limit weight can be written as

$$W_t = \begin{pmatrix} (1-\alpha_k-\alpha_\ell)D_2'\Omega^{-1}\otimes\Omega^{-1}D_2 & 0 \\ 0 & D_2'\Phi_t^{-1}D_2 \end{pmatrix} \qquad \Phi_t = \alpha_k\Omega\otimes\Omega+\Omega\otimes tmm'+tmm'\otimes\Omega$$

By Lemma C.2, $W_1 = 2\Delta^{-1}$, and

$$\Phi_t^{-1} = (\Omega^{-1}\otimes\Omega^{-1})\left[\frac{1}{\alpha_k}\left(\Omega-\frac{tmm'}{\alpha_k+t\lambda}\right)\otimes\left(\Omega-\frac{tmm'}{\alpha_k+t\lambda}\right)+\frac{t^2(mm')\otimes(mm')}{(\alpha_k+2t\lambda)(\alpha_k+t\lambda)^2}\right](\Omega^{-1}\otimes\Omega^{-1})$$

A necessary and sufficient condition for optimality is that for some matrix C_t (Newey and McFadden, 1994, Section 5.2)

$$G'W_t = C_tG'\Delta^{-1} \tag{C.32}$$

To prove the Proposition, we therefore need to prove this equality. We have:

$$W_tG = \begin{pmatrix} 0 & (1-\alpha_k-\alpha_\ell)D_2'\Omega^{-1}\otimes\Omega^{-1}D_2 \\ D_2'\Phi_t^{-1}(m\otimes M+M\otimes m) & \alpha_kD_2'\Phi_t^{-1}D_2 \end{pmatrix}$$

$$\Delta^{-1}G = \frac{1}{2}\begin{pmatrix} 0 & (1-\alpha_k-\alpha_\ell)D_2'\Omega^{-1}\otimes\Omega^{-1}D_2 \\ D_2'\Phi_1^{-1}(m\otimes M+M\otimes m) & \alpha_kD_2'\Phi_1^{-1}D_2 \end{pmatrix}$$

It therefore follows that the equality (C.32) holds with

$$C_t = \begin{pmatrix} \frac{2}{\alpha_k+t\lambda}\left((\alpha_k+\lambda)I+\frac{\alpha_k(1-t)}{\alpha+2t\lambda}M'\Omega^{-1}mm'M'^{-1}\right) & 0 \\ \frac{\alpha_k(1-t)}{\alpha_k+t\lambda}D_2'\left(\Omega^{-1}m\otimes M'^{-1}+M'^{-1}\otimes\Omega^{-1}m-\frac{2t}{(\alpha_k+2t\lambda)}\Omega^{-1}m\otimes\Omega^{-1}mm'M'^{-1}\right) & 2I_3 \end{pmatrix} \qquad \square$$

Proof of Lemma 3.1. Part (i) of the Lemma follows from Lemma C.3. Next, it it follows from Lemma A.5

in Anatolyev (2011) that:

$$\sum_i \hat{u}_i \otimes \hat{u}_i \otimes \hat{u}_i' = \sum_{ij} (M)_{ij}^3 \mathbb{E}[u_i \otimes u_i \otimes u_i'] + O(1)$$

$$\sum_i \hat{u}_i \hat{u}_i' \otimes \hat{u}_i \hat{u}_i' = \sum_{ij} (M)_{ij}^4 \mathbb{E}[u_i u_i' \otimes u_i u_i'] + \left[\delta_M - \sum_{ij} M_{ij}^4 \right] ((I_4 + K_{2,2})\Omega \otimes \Omega + \operatorname{vec}(\Omega) \operatorname{vec}(\Omega)') + O(1)$$

Part (ii) then follows. □

Proof of Lemma 3.2. The objective function evaluates as:

$$\mathcal{Q}_n(\beta, \Lambda_{22}; \hat{W}_{\text{RE}}) = \operatorname{tr}((TS^{-1} - (k_n/n) I_2)^2) - 2\Lambda_{22} a' S^{-1} T S^{-1} a + 2(k_n/n)\Lambda_{22} a' S^{-1} a + \Lambda_{22}^2 (a' S^{-1} a)^2$$

Setting derivative wrt Λ_{22} to zero:

$$\hat{\Lambda}_{22}(\beta) = \frac{Q_{\mathcal{T}}(\beta, S) - k_n/n}{a' S^{-1} a}$$

Therefore, the objective function with Λ_{22} concentrated out is given by:

$$\mathcal{Q}_n(\beta, \hat{\Lambda}_{22}(\beta)) = \operatorname{tr}((TS^{-1} - (k_n/n) I_2)^2) - (Q_{\mathcal{T}}(\beta, S) - k_n/n)^2$$

which is maximized at $\max_\beta Q_{\mathcal{T}}(\beta, S)$, since by the identity (C.1a), $\operatorname{tr}(S^{-1}T) > 2k_n/n$ implies $Q_{\mathcal{T}}(\beta, S) - k_n/n > k_n/n Q_{\mathcal{S}}(\beta, S)$, and $\min_\beta Q_{\mathcal{S}}(\beta, S) = \min_\beta Q_{\mathcal{T}}(\beta, S) = m_{\min}$. Hence, $\hat{\beta}_{\text{MD}} = \hat{\beta}_{\text{LIML}}$. Since $\hat{a}_{\text{RE}}' \hat{\Omega}_{\text{RE}}^{-1} \hat{a}_{\text{RE}} = \hat{a}_{\text{RE}}' S^{-1} \hat{a}_{\text{RE}}$ by Equation (C.30), it follows that

$$\hat{\Lambda}_{\text{MD},22} = \frac{Q_{\mathcal{T}}(\hat{\beta}_{\text{RE}}, S) - k_n/n}{\hat{a}_{\text{RE}}' S^{-1} \hat{a}_{\text{RE}}} = \frac{m_{\max} - k_n/n}{\hat{a}_{\text{RE}}' S^{-1} \hat{a}_{\text{RE}}} = \hat{\Lambda}_{\text{RE},22}$$ □

Proof of Proposition 3.6. Denote the parameters by $\theta = (\beta, \Lambda_{11}, \Lambda_{22}, \Omega)$. The log-likelihood can be written as:

$$\log \mathcal{L}_{\text{URE},n}(\theta) = -\frac{1}{2} \left((n - k_n - \ell_n)(\log |\Omega| + \operatorname{tr}(\Omega^{-1} S)) + k_n \log D(\theta) + \frac{n R(\theta)}{D(\theta)} \right) \tag{C.33}$$

where

$$R(\theta) = |\Omega| \operatorname{tr}(\Omega^{-1} T) + \frac{n}{k_n} \Lambda_{22} b' T b + \frac{n}{k_n} \Lambda_{11} T_{22}$$

The strategy for the rest of the proof is as follows. I first maximise the likelihood without imposing the constraint $\Lambda_{11} \geq 0$, yielding an estimator $\tilde{\theta}$. I then check whether the constraint binds. If $\tilde{\Lambda}_{11} \geq 0$, then it doesn't bind and $\hat{\theta}_{\text{URE}} = \tilde{\theta}$. If $\tilde{\Lambda}_{11} < 0$, then the constraint binds, and $\hat{\Lambda}_{\text{URE},11} = 0$. Moreover, since $\log \mathcal{L}_{\text{URE},n}(\beta, 0, \Lambda_{22}, \Omega) = \log \mathcal{L}_{\text{RE},n}(\beta, \Lambda_{22}, \Omega)$, the remaining estimators are equal to the RE estimators given by Proposition 3.2.

The score equations for the unconstrained likelihood are given by:

$$\mathcal{S}_\beta(\theta) = \frac{n\Lambda_{22}(n/k_n)}{D(\theta)} e_2' \left[T - \Omega \left(\frac{R(\theta)}{D(\theta)} - \frac{k_n}{n} \right) \right] b \tag{C.34a}$$

$$\mathcal{S}_{\Lambda_{11}}(\theta) = -\frac{n^2/k_n}{2D(\theta)} \left[T_{22} - \left(\Omega_{22} + \frac{n}{k_n}\Lambda_{22} \right) \left(\frac{R(\theta)}{D(\theta)} - \frac{k_n}{n} \right) \right] \tag{C.34b}$$

$$\mathcal{S}_{\Lambda_{22}}(\theta) = -\frac{n^2/k_n}{2D(\theta)} \left[b'Tb - \left(b'\Omega b + \frac{n}{k_n}\Lambda_{11} \right) \left(\frac{R(\theta)}{D(\theta)} - \frac{k_n}{n} \right) \right] \tag{C.34c}$$

$$\begin{aligned} \mathcal{S}_\Omega(\theta) = &-\frac{n - k_n - \ell_n}{2} \left(\Omega^{-1} - \Omega^{-1} S \Omega^{-1} \right) - \frac{n}{2D(\theta)} \begin{pmatrix} T_{22} & -T_{12} \\ -T_{12} & T_{11} \end{pmatrix} \\ &+ \frac{n}{2D(\theta)} \left(|\Omega| \Omega^{-1} + \frac{n}{k_n}\Lambda_{22} bb' + \frac{n}{k_n}\Lambda_{11} e_2 e_2' \right) \left(\frac{R(\theta)}{D(\theta)} - \frac{k_n}{n} \right) \end{aligned} \tag{C.34d}$$

where I use (C.1) in the derivation of $S_{\text{URE},n,\Omega}$. Setting these equations to zero yields the unrestricted ML estimators:

$$\tilde{\beta} = \frac{T_{12} - (k_n/n)S_{12}}{T_{22} - (k_n/n)S_{22}} \qquad \tilde{\Lambda}_{22} = T_{22} - (k_n/n)S_{22}$$

$$\tilde{\Lambda}_{11} = \tilde{b} \left(T - \frac{k_n}{n} S \right) \tilde{b} \qquad \tilde{\Omega} = S$$

Next, note that

$$\begin{aligned} \tilde{\Lambda}_{11} &= \frac{|T - (k_n/n)S|}{\tilde{\Lambda}_{22}} = \frac{|S| \left| S^{-1}T - (k_n/n) I_2 \right|}{\tilde{\Lambda}_{22}} \\ &= \frac{|S|}{\tilde{\Lambda}_{22}} (m_{\max} - k_n/n)(m_{\min} - k_n/n) \end{aligned}$$

Hence, under since $m_{\max} > k_n/n$ by assumption, $\hat{\Lambda}_{\text{URE},11} \geq 0$ if and only if

$$m_{\min} \geq k_n/n$$

Otherwise, the constraint $\Lambda_{11} \geq 0$ binds, in which case $\hat{\theta}_{\text{URE}} = \hat{\theta}_{\text{RE}}$.

Next, to establish part (ii), note that under MI and ODE:

$$\begin{aligned} S &\xrightarrow{p} \Omega \\ T &\xrightarrow{p} \begin{pmatrix} \Lambda_{11} + \Lambda_{22}\beta^2 & \Lambda_{22}\beta \\ \Lambda_{22}\beta & \Lambda_{22} \end{pmatrix} + \alpha_k \Omega \end{aligned}$$

The consistency of the estimators follows by simple algebra. □

Proof of Proposition 3.7. Let $\theta = (\Lambda_{11}, \Lambda_{22}, \beta)$. The derivative of the objective function is given by:

$$\begin{aligned}
\frac{\partial \mathcal{Q}_n}{\partial \Lambda_{11}} &= \operatorname{vec}(e_1 e_1)'(S \otimes S)^{-1} \operatorname{vec}\left(T - (k_n/n)S - M\right) \\
\frac{\partial \mathcal{Q}_n}{\partial \Lambda_{22}} &= \operatorname{vec}(aa')'(S \otimes S)^{-1} \operatorname{vec}\left(T - (k_n/n)S - M\right) \\
\frac{\partial \mathcal{Q}_n}{\partial \beta} &= \Lambda_{22} \operatorname{vec}(ae_1' + e_1 a')(S \otimes S)^{-1} \operatorname{vec}\left(T - (k_n/n)S - M\right)
\end{aligned}$$

Using properties of the vec operator, this first-order condition can be rewritten as:

$$\frac{\partial \mathcal{Q}_n}{\partial \Lambda_{11}} = e_1' S^{-1} \left(T - (k_n/n)S - \Lambda_{11} e_1 e_1' - \Lambda_{22} aa'\right) S^{-1} e_1 \tag{C.35a}$$

$$\frac{\partial \mathcal{Q}_n}{\partial \Lambda_{22}} = a' S^{-1} \left(T - (k_n/n)S - \Lambda_{11} e_1 e_1' - \Lambda_{22} aa'\right) S^{-1} a \tag{C.35b}$$

$$\frac{\partial \mathcal{Q}_n}{\partial \beta} = 2e_1' S^{-1} \left(T - (k_n/n)S - \Lambda_{11} e_1 e_1' - \Lambda_{22} aa'\right) S^{-1} a \tag{C.35c}$$

Denote the minimum distance estimator based on the objective function by $(\hat{\Lambda}_{11}, \hat{\Lambda}_{22}, \hat{\beta})$. If the restriction $\Lambda_{11} = 0$ is imposed, then (C.35b) implies:

$$\hat{\Lambda}_{22} = \frac{\hat{a}' S^{-1} \left(T - (k_n/n)S\right) S^{-1} \hat{a}}{(\hat{a}' S^{-1} \hat{a})^2} = \frac{1}{\hat{a}' S^{-1} \hat{a}} \left[Q_{\mathcal{T}}(S^{-1}, \hat{\beta}) - k_n/n\right]$$

Since $\hat{\beta}$ needs to minimize the objective function, it follows that:

$$\begin{aligned}
\hat{\beta} &= \operatorname*{argmin}_{\beta} \operatorname{tr}\left(\left(S^{-1} T S^{-1} - \alpha_k S^{-1} - \frac{Q_{\mathcal{T}}(S^{-1}, \beta) - \alpha_k}{a' S^{-1} a} S^{-1} aa' S^{-1}\right)\left(T - \alpha_k S - \frac{Q_{\mathcal{T}}(S^{-1}, \beta) - \alpha_k}{a' S^{-1} a} aa'\right)\right) \\
&= \operatorname{argmin} - \left(Q_{\mathcal{T}}(S^{-1}, \beta) - \alpha_k\right)^2
\end{aligned}$$

This implies that $\hat{\beta} = \hat{\beta}_{\text{LIML}}$ since by the identity (C.1a), $\operatorname{tr}(S^{-1}T) > 2k_n/n$ implies $Q_{\mathcal{T}}(\beta, S) - k_n/n > k_n/n Q_{\mathcal{S}}(\beta, S)$, and $\min_\beta Q_{\mathcal{S}}(\beta, S) = \min_\beta Q_{\mathcal{T}}(\beta, S) = m_{\min}$. Hence:

$$\hat{\Lambda}_{22} = \frac{1}{\hat{a}_{\text{LIML}}' S^{-1} \hat{a}_{\text{LIML}}} \left[Q_{\mathcal{T}}(S^{-1}, \hat{\beta}_{\text{LIML}}) - k_n/n\right] = \frac{\hat{\lambda}_{\text{RE}}}{\hat{a}_{\text{LIML}}' S^{-1} \hat{a}_{\text{LIML}}}$$

It follows from Equation (C.30) in the proof of Proposition 3.4 that $\hat{a}_{\text{LIML}}' S^{-1} \hat{a}_{\text{LIML}} = \hat{a}_{\text{RE}}' \hat{\Omega}_{\text{RE}}^{-1} \hat{a}_{\text{RE}}$, which proves Part (i). Next, setting the first-order conditions (C.35) to zero yields:

$$(\hat{\beta}, \hat{\Lambda}_{11}, \hat{\Lambda}_{22}) = (\hat{\beta}_{\text{MBTSLS}}, \hat{b}_{\text{MBTSLS}}' \left(T - (k_n/n)\hat{\Omega}_{\text{URE}}\right) \hat{b}_{\text{MBTSLS}}, T_{22} - (k_n/n)S_{22})$$

which proves Part (iii). Finally, the unrestricted estimator of Λ_{11} is positive if and only if $m_{\min} \geq k_n/n$ by arguments identical to those in the proof of Proposition 3.6, in which case the estimator in Part (ii) equals the estimator in Part (iii). Otherwise, the objective function is minimized at the boundary, and the estimator equals the estimator in Part (i). □

Proof of Proposition 3.8. Denote the true parameter values by $\theta = (\Lambda_{11}, \Lambda_{22}, \beta)$, where $\Lambda_{11} = 0$. I first verify assumptions GMM1, MD2, and GMM3–5 in Andrews (2002). Consistency of $\hat{\theta}$, and hence assumption GMM1 follows from Proposition 3.6. Assumption MD2 follows from $T - \alpha_k S \xrightarrow{p} \Xi$, $\hat{W}^{1/2}_{\text{SIMP,RE}} = (S^{-1/2} \otimes S^{-1/2}) D_2 \xrightarrow{p} (\Omega^{-1/2} \otimes \Omega^{-1/2}) D_2 \equiv A$, and the Taylor expansion $\operatorname{vech} \Xi(\tilde{\theta}) = \operatorname{vech} \Xi(\theta) + G(\tilde{\theta} - \theta) + o(\|\tilde{\theta} - \theta\|)$ where, the derivative of the moment condition, G, is given by:

$$G = L_2 \begin{pmatrix} e_1 \otimes e_1 & a \otimes a & \Lambda_{22}(a \otimes e_1 + a \otimes a) \end{pmatrix} = L_2(\Gamma \otimes \Gamma)\tilde{D} \qquad \tilde{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & \Lambda_{22} \\ 0 & 0 & \Lambda_{22} \\ 0 & 1 & 0 \end{pmatrix}$$

Assumption GMM3 follows from an application of Corollary C.1, which yields $\sqrt{n} \operatorname{vech}(T - (k_n/n)S - \Xi(\theta_0)) \Rightarrow \mathcal{N}(0, \mathcal{V})$, where

$$\mathcal{V} = 2L_2 N_2 \left(\tau \Omega \otimes \Omega + \Omega \otimes a a' \Lambda_{22} + a a' \Lambda_{22} \otimes \Omega \right) L_2' \qquad \tau = \frac{\alpha_k (1 - \alpha_\ell)}{1 - \alpha_k - \alpha_\ell}$$

Assumption GMM4 also holds since the shifted parameter space $\Theta - \theta$ is locally equal to $\mathbb{R}_+ \times \mathbb{R}^2$. Finally, since $\mathbb{R}_+ \times \mathbb{R}^2$ is convex, Assumption GMM5 also holds. Therefore, by Theorem 1 in Andrews (2002):

$$\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \tilde{\psi} = \underset{\psi \in \mathbb{R}_+ \times \mathbb{R}^2}{\operatorname{argmin}} \, (\psi - \tilde{Z})' \mathcal{I} (\psi - \tilde{Z}) \qquad \tilde{Z} \sim \mathcal{I}^{-1} G' W \mathcal{N}(0, \mathcal{V}) \tag{C.36}$$

where $W = D_2'(\Omega^{-1} \otimes \Omega^{-1}) D_2 = A'A = \operatorname{plim} \hat{W}_{\text{SIML,RE}}$, and

$$\mathcal{I} \equiv G' W G = G' D_2' (\Omega^{-1} \otimes \Omega^{-1}) D_2 G$$

The minimization problem in (C.36) solves as (see Theorem 2 and Section 3.8 in Andrews (2002) for details):

$$\tilde{\psi}_1 = \max(0, \tilde{Z}_1) \qquad \tilde{\psi}_{2:3} = \tilde{Z}_{2:3} + \mathcal{I}^{-1}_{2:3,2:3} \mathcal{I}_{2:3,1} \min(\tilde{Z}_1, 0)$$

Hence:

$$\sqrt{n}(\hat{\beta}_{\text{URE}} - \beta) \Rightarrow \tilde{Z}_3 + \frac{\mathcal{I}_{22}\mathcal{I}_{13} - \mathcal{I}_{23}\mathcal{I}_{12}}{\mathcal{I}_{22}\mathcal{I}_{33} - \mathcal{I}_{23}^2} \min(\tilde{Z}_1, 0) \sim \tilde{Z}_3 - \frac{\Sigma_{12}}{\Lambda_{22}\Sigma_{11}} \min(\tilde{Z}_1, 0) \tag{C.37}$$

The remainder of the proof simplifies this expression by explicitly deriving the expression $\operatorname{var}(\tilde{Z})$. Define

$$\tilde{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \Lambda_{22}^{-1} & 0 & 0 \end{pmatrix}$$

Because of the similarity with duplication an elimination matrices, a result similar to Lemma 4.4(vi) in Magnus and Neudecker (1980) applies to $\tilde{D}$ and $\mathcal{L}$, namely that for any invertible 2×2 matrix A, $(\tilde{D}'(A \otimes A)\tilde{D})^{-1} = \tilde{L} N_2 (A^{-1} \otimes A^{-1}) N_2 \tilde{L}'$. Applying this result to $\mathcal{I}$ combined with the equality from Equation (3.4) $\Omega = \Gamma \Sigma \Gamma'$, so

that $\Sigma^{-1} = \Gamma'\Omega^{-1}\Gamma$, we get:

$$\mathcal{I} = \tilde{D}'(\Sigma^{-1} \otimes \Sigma^{-1})\tilde{D} \qquad\qquad \mathcal{I}^{-1} = \tilde{L}N_2(\Sigma \otimes \Sigma)N_2\tilde{L}' \tag{C.38}$$

and the equalities (C.3a), some algebra yields:

$$\text{var}(\tilde{Z}) = 2\tilde{L}N\left(\tau\Sigma \otimes \Sigma + \Sigma \otimes e_2e_2'\Lambda_{22} + e_2e_2'\Lambda_{22} \otimes \Sigma\right)N\tilde{L}' \tag{C.39}$$

Using the representation

$$N\tilde{L}' = \begin{pmatrix} e_1 \otimes e_1 & e_2 \otimes e_2 & \frac{1}{2\Lambda_{22}}\left(e_1 \otimes e_2 + e_2 \otimes e_1\right) \end{pmatrix}$$

we can write (C.39) as:

$$\text{var}(\tilde{Z}) = \begin{pmatrix} 2\tau\Sigma_{11}^2 & 2\tau\Sigma_{12}^2 & \frac{2\tau}{\Lambda_{22}}\Sigma_{11}\Sigma_{12} \\ 2\tau\Sigma_{12}^2 & 2\tau\Sigma_{22}^2 + 4\Sigma_{22}\Lambda_{22} & \frac{2\tau\Sigma_{12}\Sigma_{22}}{\Lambda_{22}} + 2\Sigma_{12} \\ \frac{2\tau}{\Lambda_{22}}\Sigma_{11}\Sigma_{12} & \frac{2\tau\Sigma_{12}\Sigma_{22}}{\Lambda_{22}} + 2\Sigma_{12} & \frac{\tau}{\Lambda_{22}^2}[\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2] + \Sigma_{11}/\Lambda_{22} \end{pmatrix}$$

Hence:

$$\sqrt{n}(\hat{\beta}_{\text{URE}} - \beta) \Rightarrow \tilde{Z}_3 - \frac{\Sigma_{12}}{\Lambda_{22}\Sigma_{11}}\min(\tilde{Z}_1, 0) \quad \begin{pmatrix} \tilde{Z}_1 \\ \tilde{Z}_3 \end{pmatrix} \sim \mathcal{N}_2\left(0, \begin{pmatrix} 2\tau\Sigma_{11}^2 & \frac{2\tau}{\Lambda_{22}}\Sigma_{11}\Sigma_{12} \\ \frac{2\tau}{\Lambda_{22}}\Sigma_{11}\Sigma_{12} & \frac{\tau}{\Lambda_{22}^2}[\Sigma_{11}\Sigma_{22} + \Sigma_{12}^2] + \frac{\Sigma_{11}}{\Lambda_{22}} \end{pmatrix}\right) \tag{C.40}$$

Using Cholesky decomposition, this can be written as:

$$\tilde{Z} \sim \begin{pmatrix} \sqrt{2\tau}\Sigma_{11} & 0 \\ \frac{\sqrt{2\tau}}{\Lambda_{22}}\Sigma_{12} & \sqrt{\frac{\tau}{\Lambda_{22}^2}[\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2] + \Sigma_{11}/\Lambda_{22}} \end{pmatrix}\mathcal{Z} \qquad \mathcal{Z} \sim \mathcal{N}_2(0, I_2)$$

which yields the result. □

Proof of Lemma 3.3. We have:

$$\begin{aligned}
\sqrt{n}(m_{\min} - \alpha_k) &= \sqrt{n}\frac{\hat{b}_{\text{LIML}}'(T - \alpha_k S)\hat{b}_{\text{LIML}}}{\hat{b}_{\text{LIML}}'S\hat{b}_{\text{LIML}}} \\
&= \frac{\hat{b}_{\text{LIML}}'\sqrt{n}(T - \alpha_k S - \lambda_n aa'/(a'\Omega^{-1}a))\hat{b}_{\text{LIML}}}{\hat{b}_{\text{LIML}}'S\hat{b}_{\text{LIML}}} + \frac{\sqrt{n}\lambda_n(a'\hat{b}_{\text{LIML}})^2}{(a'\Omega^{-1}a)\hat{b}_{\text{LIML}}'S\hat{b}_{\text{LIML}}} \\
&= \frac{\sqrt{n}(\hat{b}_{\text{LIML}} \otimes \hat{b}_{\text{LIML}})'\,\text{vec}\left(T - \alpha_k S - \lambda_n aa'/(a'\Omega^{-1}a)\right)}{\hat{b}_{\text{LIML}}'S\hat{b}_{\text{LIML}}} + \frac{\sqrt{n}\lambda_n(\hat{b}_{\text{LIML}} - \beta)^2}{(a'\Omega^{-1}a)\hat{b}_{\text{LIML}}'S\hat{b}_{\text{LIML}}} \\
&= \frac{\sqrt{n}(\hat{b}_{\text{LIML}} \otimes \hat{b}_{\text{LIML}})'\,\text{vec}\left(T - \alpha_k S - \lambda_n aa'/(a'\Omega^{-1}a)\right)}{\hat{b}_{\text{LIML}}'S\hat{b}_{\text{LIML}}} + o_p(1)
\end{aligned}$$

where the first line follows from the identity $m_{\min} = Q_{\mathcal{S}}(S, \hat{\beta}_{\text{LIML}})$, the second line follows by algebra, the third line follows from (C.2), and the fourth line follows from $\hat{\beta}_{\text{LIM}} - \beta = O_p(n^{-1/2})$. Using Lemma C.1, consistency of $\hat{\beta}_{\text{LIML}}$, the continuous mapping theorem and (C.3a), we obtain:

$$\sqrt{n}(\hat{b}_{\text{LIML}} \otimes \hat{b}_{\text{LIML}})'\,\text{vec}\left(T - \alpha_k S - \lambda_n aa'/(a'\Omega^{-1}a)\right) \Rightarrow \mathcal{N}\left(0, 2\tau(b'\Omega b)^2\right) \tag{C.41}$$

where $\tau = \alpha_k(1-\alpha_\ell)/(1-\alpha_k-\alpha_\ell)$. Combining these results, we get:

$$\sqrt{n}(m_{\min} - \alpha_k) \Rightarrow \mathcal{N}(0, 2\tau)$$

The results for $\hat{J}_{\text{S}}$ and $\hat{J}_{\text{AR}}$ follow by the Delta method. To prove the remainder of the lemma, I use the approximation from Peiser (1943) (see also (Anatolyev and Gospodinov, 2011)) that as $k \to \infty$,

$$q_{\text{ns}}^{\chi^2_k} = k + \Phi^{-1}(1-\text{ns})\sqrt{2k} + O(1)$$

Therefore:

$$\begin{aligned}
\mathbb{P}\left(n\hat{J}_{\text{CD}} \geq q_{\text{ns}}^{\chi^2_{k_n-1}}\right) &= \mathbb{P}\left(\sqrt{n}\hat{J}_{\text{CD}} \geq k_n/\sqrt{n} + \Phi^{-1}(1-\text{ns})\sqrt{2\alpha_k} + o(1)\right) \\
&= \mathbb{P}\left(\sqrt{n}(\hat{J}_{\text{CD}} - \alpha_k)/\sqrt{2\tau} \geq \Phi^{-1}(1-\text{ns})\sqrt{\alpha_k/\tau} + o(1)\right) \\
&= \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \Phi^{-1}(1-\text{ns})\sqrt{\alpha_k/\tau} + o(1)\right) = 1 - \Phi\left(\Phi^{-1}(1-\text{ns})\sqrt{\alpha_k/\tau}\right) + o(1) \\
&\to \Phi\left(\Phi^{-1}(\text{ns})\sqrt{\alpha_k/\tau}\right)
\end{aligned}$$

Secondly,

$$\begin{aligned}
\mathbb{P}\left(n\hat{J}_{\text{S}} \geq q_{\text{ns}}^{\chi^2_{k_n-1}}\right) &= \mathbb{P}\left(\sqrt{n}\hat{J}_{\text{S}} \geq k_n/\sqrt{n} + \Phi^{-1}(1-\text{ns})\sqrt{2\alpha_k} + o(1)\right) \\
&= \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \left(\frac{(1-\alpha_\ell)^3}{2\alpha_k(1-\alpha_k-\alpha_\ell)}\right)^{1/2}\left(\Phi^{-1}(1-\text{ns})\sqrt{2\alpha_k} - \sqrt{n}\frac{\alpha_k\alpha_\ell}{1-\alpha_\ell}\right) + o(1)\right)
\end{aligned}$$

Now, if $\alpha_k > 0$, then the right-hand side converges to $-\infty$, so that the rejection probability converges to one. If $\alpha_k = 0$, then

$$\mathbb{P}\left(n\hat{J}_{\text{S}} \geq q_{\text{ns}}^{\chi^2_{k_n-1}}\right) = \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \frac{\Phi^{-1}(1-\text{ns})}{(1-\alpha_k)^{1/2}} + o(1)\right) \to \Phi\left(\frac{\Phi^{-1}(\text{ns})}{(1-\alpha_k)^{1/2}}\right)$$

Thirdly,

$$\begin{aligned}
&\mathbb{P}\left(n\hat{J}_{\text{AR}} \geq q_{\text{ns}}^{\chi^2_{k_n-1}}\right) \\
&= \mathbb{P}\left(\sqrt{n}\hat{J}_{\text{AR}} \geq k_n/\sqrt{n} + \Phi^{-1}(1-\text{ns})\sqrt{2\alpha_k} + o(1)\right) \\
&= \mathbb{P}\left(\mathcal{N}(0,1) + o_p(1) \geq \frac{1-\alpha_\ell}{\sqrt{2\tau}}\left(\sqrt{n}\left[\alpha_k - \log\left(\frac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}\right)\right] + \Phi^{-1}(1-\text{ns})\sqrt{2\alpha_k} + o(1)\right)\right)
\end{aligned}$$

Now, since $\alpha_k < -\log(1-\alpha_k)$,

$$\alpha_k - \log\left(\frac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}\right) < \log\left(\frac{1}{1-\alpha_k}\right) - \log\left(\frac{1-\alpha_\ell}{1-\alpha_k-\alpha_\ell}\right) = \log\left(\frac{1-\alpha_k-\alpha_\ell}{(1-\alpha_k)(1-\alpha_\ell)}\right) < \log(1) = 0$$

so that the right-hand side of the expression converges to $-\infty$, and the rejection probability converges to 1. □

Proof of Lemma 3.4. Denote the parameters of the model by $\theta = (\beta, \Lambda_{11}, \Lambda_{22}, \Omega)$. Suppose that $m_{\min} \geq$

k_n/n, otherwise both test statistics are trivially equal to zero. I first derive the expression for the generalized likelihood ratio test statistic, which is given by:

$$\hat{\zeta}_{\text{LR}} = 2(\log \mathcal{L}_{\text{URE},n}(\hat{\theta}_{\text{URE}}) - \log \mathcal{L}_{\text{URE},n}(\hat{\theta}_{\text{RE}}))$$

where $\mathcal{L}_{\text{URE},n}(\theta)$ is given by Equation (C.33). First consider $\log \mathcal{L}_{\text{URE},n}(\hat{\theta}_{\text{URE}})$. Since $D(\hat{\theta}_{\text{URE}}) = |nT/k_n|$, plugging in $\hat{\theta}_{\text{URE}}$ in place of the remaining parameters yields:

$$\log \mathcal{L}_{\text{URE}}(\hat{\theta}_{\text{URE}}) = -\frac{1}{2}\left((n-k_n-\ell_n)\log|S| + 2(n-\ell_n) + k_n \log|nT/k_n|\right)$$

Let:

$$R = \frac{\hat{\lambda}_{\text{RE}}}{\hat{\lambda}_{\text{RE}} + k_n/n} \frac{\hat{a}'_{\text{RE}}(Y'_\perp Y_\perp)^{-1}\hat{a}_{\text{RE}}}{\hat{a}'_{\text{RE}}(nT)^{-1}\hat{a}_{\text{RE}}}$$

Then we can write:

$$\text{tr}(\hat{\Omega}_{\text{RE}}^{-1} Y'_\perp Y_\perp) = (n-\ell_n)\left(2 + \frac{R}{1-R}\right) \qquad |\hat{\Omega}_{\text{RE}}| = \frac{1}{(n-\ell_n)^2}\left|Y'_\perp Y_\perp\right|(1-R)$$

Using Equations (C.1), the expression for R can be simplified to:

$$R = 1 - (n-\ell_n)\frac{(nm_{\min} + n - k_n - \ell_n)\,|S|}{Y'_\perp Y_\perp}$$

Hence, using Equations (C.1) again:

$$|\hat{\Omega}_{\text{RE}}| = \frac{n-k_n-\ell_n+nm_{\min}}{n-\ell_n}|S|$$
$$\text{tr}(\hat{\Omega}_{\text{RE}}^{-1} Y'_\perp Y_\perp) = 2(n-\ell_n) + nm_{\max} - k_n$$

We therefore have:

$$\log \mathcal{L}_{\text{RE}}(\hat{\theta}_{\text{RE}}) =$$
$$-\frac{1}{2}\left((n-\ell_n)\log\frac{n-k_n-\ell_n+nm_{\min}}{(n-\ell_n)(n-k_n-\ell_n)} + (n-\ell_n)\log((n-k_n-\ell_n)|S|) + k_n\log\left(\frac{nm_{\max}}{k_n}\right)\right) - (n-\ell_n)$$

So that:

$$\hat{\zeta}_{\text{LR}} = (n-\ell_n)\log\left(\frac{n-k_n-\ell_n+nm_{\min}}{(n-\ell_n)}\right) - k_n\log\left(\frac{nm_{\min}}{k_n}\right)$$

which yields the result.

Now consider the minimum distance objective function. At $(\hat{\beta}_{\text{RE}}, \hat{\lambda}_{\text{RE}}, \hat{\Omega}_{\text{RE}})$, which I denote by $(\hat{\beta}, \hat{\lambda}, \hat{\Omega})$ to reduce clutter, the objective function evaluates as:

$$\mathcal{Q}_n(\hat{\beta}, \hat{\lambda}, \hat{\Omega}) = \frac{n-k_n-\ell_n}{n}\text{tr}((I_2 - S^{-1}\hat{\Omega})^2) + (k_n/n)\,\text{tr}(((n/k_n)S^{-1}T - S^{-1}\hat{\Omega})^2) - \frac{n}{k_n}\lambda^2$$

where I use the identities $\hat{a}'\hat{\Omega}^{-1}\hat{a} = \hat{a}'S^{-1}\hat{a}$ (see Equation (C.30)) and $\hat{a}S^{-1}\hat{\Omega}S^{-1}\hat{a} = \hat{a}S^{-1}\hat{a}$ (which follows from the definition of $\hat{\Omega}$ and some algebra). Next, since

$$\text{tr}\left(((k_n/n)I_2 - S^{-1}T)^2\right) = (m_{\max} - k_n/n)^2 + (m_{\min} - k_n/n)^2$$

we have:

$$\begin{aligned}
\frac{k_n}{n}\,\text{tr}((n/k_n)S^{-1}T - S^{-1}\hat{\Omega})^2 &= \frac{k_n}{n}\left(\frac{n}{n-\ell_n}\right)^2\left(\frac{(n-k_n-\ell_n)^2}{k_n^2}\,\text{tr}(\frac{k_n}{n}I_2 - S^{-1}T)^2 + \hat{\lambda}^2 + \frac{2\hat{\lambda}^2(n-k_n-\ell_n)}{k_n}\right) \\
&= \frac{k_n}{n}\left(\frac{n}{n-\ell_n}\right)^2\left(\frac{(n-\ell_n)^2}{k_n^2}\hat{\lambda}^2 + \frac{(n-k_n-\ell_n)^2}{k_n^2}(m_{\min} - k_n/n)^2\right) \\
&= \frac{n}{k_n}\hat{\lambda}^2 + \frac{n(n-k_n-\ell_n)^2}{k_n(n-\ell_n)^2}(m_{\min} - k_n/n)^2 \\
\frac{n-k_n-\ell_n}{n}\text{tr}((I_2 - S^{-1}\hat{\Omega})^2) &= \frac{n-k_n-\ell_n}{n}\left(\frac{n}{n-\ell_n}\right)^2\left(\text{tr}((k_n/n)I_2 - S^{-1}T)^2 - \hat{\lambda}^2\right) \\
&= \frac{(n-k_n-\ell_n)n}{(n-\ell_n)^2}\,(m_{\min} - k_n/n)^2
\end{aligned}$$

Therefore, we obtain:

$$\mathcal{Q}_n(\hat{\beta}, \hat{\lambda}, \hat{\Omega}) = \frac{n-k_n-\ell_n}{k_n/n(n-\ell_n)}\,(m_{\min} - k_n/n)^2$$

which yields the result. □